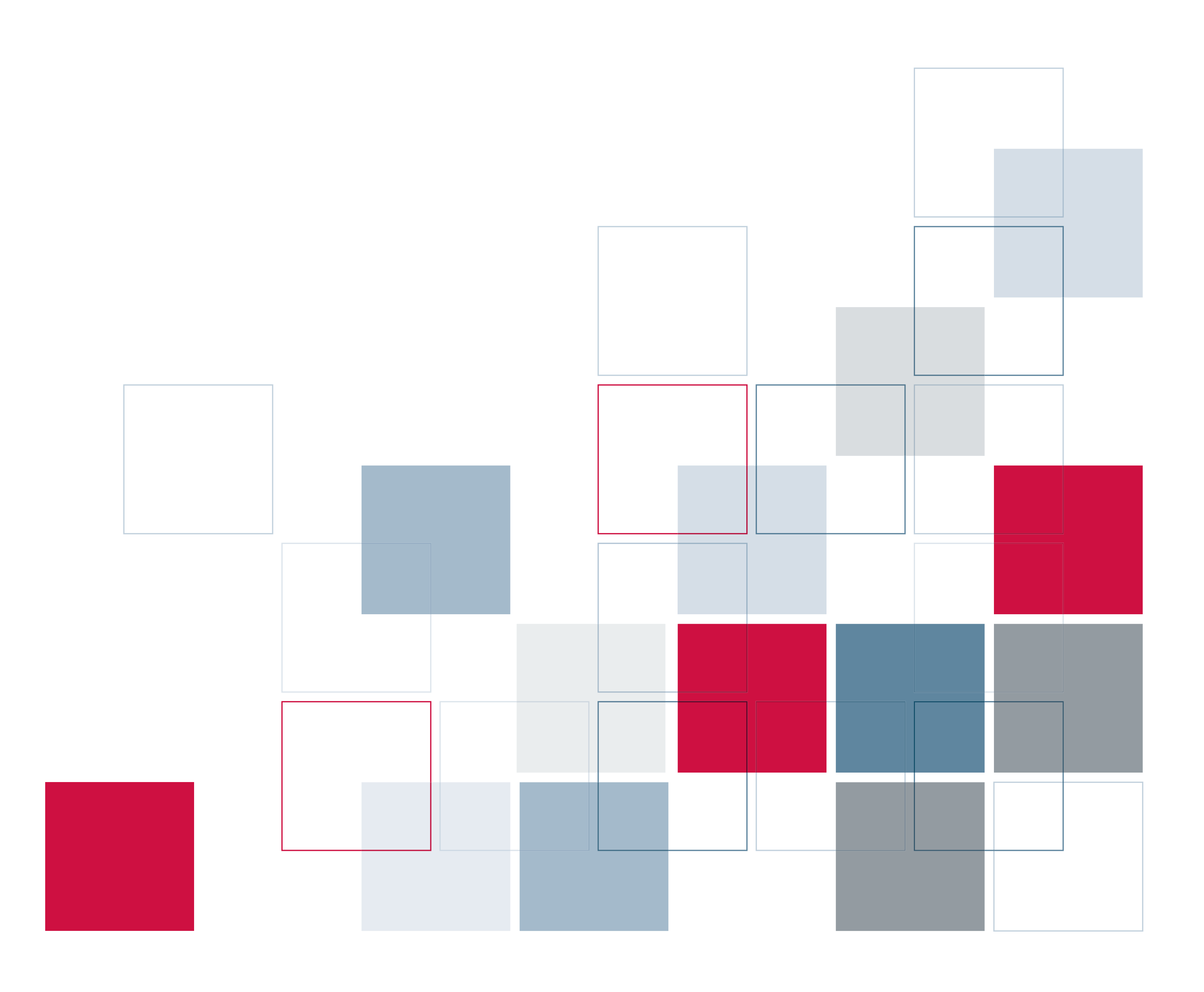
PASW® Direct Marketing (Direktmarketing) 18



Weitere Informationen zu SPSS Inc.-Software-Produkten finden Sie auf unserer Website unter der Adresse *http://www.spss.com* oder wenden Sie sich an

SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6412, USA
Tel.: (312) 651-3000
Fax: (312) 651-3668

Vorwort

PASW Statistics 18 ist ein umfassendes System zum Analysieren von Daten. Das optionale Zusatzmodul Option “Direct Marketing” (Direktmarketing) bietet die zusätzlichen Analyseverfahren, die in diesem Handbuch beschrieben sind. Die Prozeduren im Zusatzmodul Option “Direct Marketing” (Direktmarketing) müssen zusammen mit PASW Statistics 18 Core verwendet werden. Sie sind vollständig in dieses System integriert.

Installation

Zur Installation von Option “Direct Marketing” (Direktmarketing) Erweiterungsmodul führen Sie den Lizenzautorisierungsassistenten mit dem Autorisierungscode aus, den Sie von SPSS Inc. erhalten haben. Weitere Informationen finden Sie in den Installationsanweisungen im Lieferumfang von Option “Direct Marketing” (Direktmarketing) Erweiterungsmodul.

Kompatibilität

PASW Statistics kann auf vielen Computersystemen ausgeführt werden. Mindestanforderungen an das System und Empfehlungen finden Sie in den Unterlagen, die mit Ihrem System geliefert werden.

Seriennummern

Die Seriennummer des Programms dient gleichzeitig als Identifikationsnummer bei SPSS Inc.. Sie benötigen diese Seriennummer, wenn Sie sich an SPSS Inc. wenden, um Informationen zum Kundendienst, zu Zahlungen oder zu Aktualisierungen des Systems zu erhalten. Die Seriennummer wird mit dem Core-System ausgeliefert.

Kundendienst

Wenden Sie sich bei Fragen zur Lieferung oder Ihrem Kundenkonto an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Halten Sie bitte stets Ihre Seriennummer bereit.

Ausbildungsseminare

SPSS Inc. bietet öffentliche und unternehmensinterne Seminare an. Alle Seminare beinhalten auch praktische Übungen. Seminare finden in größeren Städten regelmäßig statt. Wenn Sie weitere Informationen zu diesen Seminaren wünschen, wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden.

Technischer Support

Kunden mit Wartungsvertrag können den Technischen Support in Anspruch nehmen. Kunden können sich an den Technischen Support wenden, wenn sie Hilfe bei der Arbeit mit PASW Statistics oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Informationen zum Technischen Support finden Sie auf der Website unter *http://www.spss.com* oder wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Bei einem Anruf werden Sie nach Ihrem Namen, dem Namen Ihrer Organisation und Ihrer Seriennummer gefragt.

Weitere Veröffentlichungen

Das Handbuch *SPSS Statistics Statistical Procedures Companion* von Marija Norušis wurde von Prentice Hall veröffentlicht. Eine neue Fassung dieses Buchs mit Aktualisierungen für PASW Statistics 18 ist geplant. Das Handbuch *SPSS Statistics Advanced Statistical Procedures Companion*, bei dem auch PASW Statistics 18 berücksichtigt wird, erscheint demnächst. Das Handbuch *SPSS Statistics Guide to Data Analysis* für PASW Statistics 18 wird ebenfalls derzeit erstellt. Ankündigungen für Veröffentlichungen, die ausschließlich über Prentice Hall verfügbar sind, finden Sie auf der Website unter *http://www.spss.com/estore* (wählen Sie Ihr Land aus und klicken Sie auf Books).

Inhalt

Teil I: Benutzerhandbuch

11 Responseraten nach Postleitzahlen 79

12 Kaufneigung 86

13 Kontrollpakettest 99

Anhang

A Beispieldateien 103

Index 115

Teil I: Benutzerhandbuch

Kapitel

1

Option "Direct Marketing" (Direktmarketing)

Die Option "Direktmarketing" bietet eine Reihe von Werkzeugen zur Verbesserung der Ergebnisse von Direktmarketing-Kampagnen durch die Identifizierung von Demografie-, Einkaufs- und anderen Merkmalen, die unterschiedliche Kundengruppen definieren, sowie durch Konzentration auf bestimmte Gruppen zur Maximierung positiver Responseraten.

RFM-Analyse. Dieses Verfahren identifiziert bestehende Kunden, die sehr wahrscheinlich auf ein neues Angebot antworten. Für weitere Informationen siehe RFM-Analyse in Kapitel 2 auf S. 2.

Cluster-Analyse. Hierbei handelt es sich um eine explorative Prozedur zum Ermitteln von natürlichen Gruppierungen (Clustern) innerhalb Ihrer Daten. Damit können beispielsweise verschiedene Kundengruppen auf der Basis unterschiedlicher demographischer und Kaufverhaltensmerkmale ausgemacht werden. Für weitere Informationen siehe Cluster-Analyse in Kapitel 3 auf S. 15.

Profile über potenzielle Kunden. Bei dieser Technik werden Ergebnisse aus einer früheren Kampagne oder einer Testkampagne verwendet, um beschreibende Profile zu erstellen. Diese Profile können bei zukünftigen Kampagnen für das Targeting bestimmter Gruppen von Kontakten verwendet werden. Für weitere Informationen siehe Profile über potenzielle Kunden in Kapitel 4 auf S. 20.

Responseraten nach Postleitzahlen. Bei dieser Technik werden Ergebnisse aus einer früheren Kampagne verwendet, um Responseraten nach Postleitzahlen zu berechnen. Diese Raten können bei zukünftigen Kampagnen für das Targeting bestimmter Postleitzahlbereiche verwendet werden. Für weitere Informationen siehe Responseraten nach Postleitzahlen in Kapitel 5 auf S. 26.

Kaufneigung. In diesem Verfahren werden Ergebnisse einer Testsendung oder einer früheren Kampagne verwendet, um Bewertungen zu erstellen. Die Bewertungen zeigen an, bei welchen Kontakten die Wahrscheinlichkeit einer Antwort am höchsten ist. Für weitere Informationen siehe Kaufneigung in Kapitel 6 auf S. 34.

Kontrollpakettest. Dieses Verfahren vergleicht Marketingkampagnen, um herauszufinden, ob es hinsichtlich der Effektivität signifikante Unterschiede zwischen verschiedenen Paketen oder Angeboten gibt. Für weitere Informationen siehe Kontrollpakettest in Kapitel 7 auf S. 44.

Kapitel

2

RFM-Analyse

Die RFM-Analyse (Recency – Aktualität, Frequency – Häufigkeit, Monetary – Geldwert) ist eine Technik, die verwendet wird, um bestehende Kunden zu identifizieren, die am wahrscheinlichsten auf ein neues Angebot reagieren werden. Diese Technik wird häufig im Direktmarketing eingesetzt. RFM-Analyse basiert auf der folgenden einfachen Theorie:

- Der wichtigste Faktor bei der Identifizierung von Kunden, die wahrscheinlich auf ein neues Angebot reagieren, ist **Aktualität**. Kunden, die kürzlich gekauft haben, kaufen wahrscheinlicher wieder ein, als Kunden, die weiter zurück in der Vergangenheit gekauft haben.
- Der zweitwichtigste Faktor ist **Häufigkeit**. Kunden, die in der Vergangenheit häufiger gekauft haben, kaufen wahrscheinlicher wieder ein, als Kunden, die weniger gekauft haben.
- Der drittwichtigste Faktor ist der ausgegebene Betrag, der als **Geldwert** bezeichnet wird. Kunden, die in der Vergangenheit (für alle Einkäufe insgesamt) mehr ausgegeben haben, reagieren wahrscheinlicher, als Kunden, die weniger ausgegeben haben.

Funktionsweise der RFM-Analyse

- Kunden wird basierend auf dem Datum des letzten Kaufs bzw. des Zeitintervalls seit dem letzten Kauf ein Aktualitäts-Score zugewiesen. Dieser Score basiert auf einer einfachen Einstufung von Aktualitätswerten in eine kleine Zahl von Kategorien. Wenn Sie zum Beispiel fünf Kategorien verwenden, erhalten die Kunden mit den neuesten Kaufdaten eine Aktualitätseinstufung von 5 und die mit den am weitesten zurückliegenden Kaufdaten eine Aktualitätseinstufung von 1.
- Auf ähnliche Weise wird Kunden dann eine Häufigkeitseinstufung zugewiesen, wobei höhere Werte eine höhere Kaufhäufigkeit bedeutet. In einem Einstufungsschema mit fünf Kategorien erhalten Kunden, die am häufigsten einkaufen, eine Häufigkeitseinstufung von 5.
- Schließlich werden die Kunden nach Geldwert eingestuft, wobei die höchsten Geldwerte die höchste Einstufung erhalten. In dem Beispiel mit fünf Kategorien würden die Kunden, die den höchsten Betrag aufwenden, eine Geldwerteinstufung von 5 erhalten.

Das Ergebnis sind vier Scores für jeden Kunden: der Aktualitäts-, der Häufigkeits-, der Geldwert- und der kombinierte RFM-Score, bei dem einfach die drei einzelnen Scores in einem einzigen Wert aneinandergehängt werden. Die “besten” Kunden (die am wahrscheinlichsten auf ein Angebot reagieren) sind diejenigen Kunden mit den höchsten kombinierten RFM-Scores. In einer Einstufung mit fünf Kategorien gibt es zum Beispiel insgesamt 125 mögliche, kombinierte RFM-Scores, der höchste kombinierte RFM-Score ist 555.

Erläuterung der Daten

- Wenn Datenzeilen Transaktionen darstellen (jede Zeile repräsentiert eine einzelne Transaktion und es kann mehrere Transaktionen für jeden Kunden geben), verwenden Sie RFM aus Transaktionen. Für weitere Informationen siehe RFM-Scores aus Transaktionsdaten auf S. 3.
- Wenn Datenzeilen Kunden mit Auswertungsinformationen für alle Transaktionen darstellen (mit Spalten, die Werte für den Gesamtkaufbetrag, die Gesamtzahl der Transaktionen und das letzte Transaktionsdatum enthalten), verwenden Sie RFM aus Kundendaten. Für weitere Informationen siehe RFM-Scores aus Kundendaten auf S. 5.

Abbildung 2-1
Transaktion im Vergleich zu Kundendaten

Zeilen sind Transaktionen.

ID	Gender	Date	Amount
1	Male	9/25/2005	21
2	Male	1/15/2006	297
4	Male	2/5/2006	249
4	Male	5/7/2005	172
6	Male	4/16/2005	164
6	Male	4/12/2005	286
7	Femal	[illegible]	[illegible]
9	Male	[illegible]	[illegible]
9	Male	[illegible]	[illegible]
9	Male	[illegible]	[illegible]
10	Femal	[illegible]	[illegible]
10	Femal	[illegible]	[illegible]

Zeilen sind Kunden mit Transaktionsauswertungen.

ID	Gender	Most Recent	Total Amount	Number of Purchases
1	Male	9/25/2005	21	1
2	Male	1/15/2006	297	1
4	Male	2/5/2006	421	2
6	Male	4/16/2005	450	2

RFM-Scores aus Transaktionsdaten

Erläuterung der Daten

Das Daten-Set muss Variablen enthalten, die die folgenden Informationen enthalten:

- Eine Variable oder eine Kombination von Variablen, die jeden Fall (Kunden) identifizieren
- Eine Variable mit dem Datum jeder Transaktion
- Eine Variable mit dem Geldwert jeder Transaktion

Abbildung 2-2
RFM-Transaktionsdaten

ID	Date	Amount
1	08/04/2005	129
1	10/25/2004	50
1	07/24/2004	118
1	07/24/2004	136
1	09/04/2006	52
2	09/23/2005	183
2	11/05/2004	24
2	11/10/2005	66
2	12/03/2004	77
3	06/04/2005	102
3	05/15/2005	131

Erstellen von RFM-Scores aus Transaktionsdaten

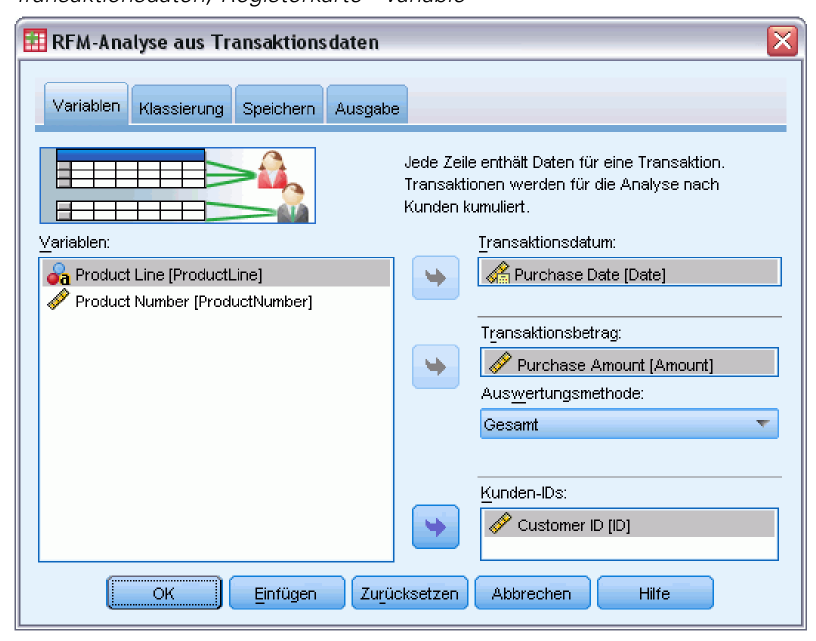
▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Hilfe bei der Erkennung meiner besten Kontakte (RFM-Analyse) und klicken Sie auf Weiter.

▶ Wählen Sie Transaktionsdaten und klicken Sie auf Weiter.

Abbildung 2-3
Transaktionsdaten, Registerkarte “Variable”



▶ Wählen Sie die Variable aus, die Transaktionsdaten enthält.

▶ Wählen Sie die Variable, die den Geldwertbetrag für jede Transaktion enthält.

▶ Wählen Sie die Methode für die Zusammenfassung der Transaktionsbeträge für jeden Kunden: Summe (Summe aller Transaktionen), Mittelwert, Median oder Maximum (höchster Transaktionsbetrag).

▶ Wählen Sie die Variable oder die Kombination von Variablen, die jeden Kunden eindeutig identifiziert. Zum Beispiel könnten Fälle durch einen eindeutigen Schlüsselcode oder eine Kombination aus Nachname und Vorname identifiziert werden.

RFM-Scores aus Kundendaten

Erläuterung der Daten

Das Daten-Set muss Variablen enthalten, die die folgenden Informationen enthalten:

- Das letzte Kaufdatum oder ein Zeitintervall seit dem letzten Kaufdatum. Dies wird zur Berechnung der Aktualitäts-Scores verwendet.
- Gesamtzahl von Käufen. Dies wird zur Berechnung der Häufigkeits-Scores verwendet.
- Gesamtgeldwertbetrag für alle Käufe. Dies wird zur Berechnung der Geldwert-Scores verwendet. In der Regel ist dies die Summe aller Käufe, könnte jedoch auch der Mittelwert (Durchschnitt), das Maximum (größter Betrag) oder eine andere Auswertungskennzahl sein.

Abbildung 2-4
RFM-Kundendaten

ID	TotalAmount	MostRecent	NumberOfPurchases
1	485.00	09/04/2006	5
2	350.00	11/10/2005	4
3	233.00	06/04/2005	2
4	936.00	08/18/2006	7
5	359.00	07/07/2006	3
6	249.00	07/16/2006	3
7	1089.00	02/15/2006	7
8	423.00	08/21/2006	4
9	689.00	08/31/2006	7
10	325.00	10/13/2005	3

Wenn Sie die RFM-Scores in ein neues Daten-Set schreiben möchten, muss das aktive Daten-Set auch eine Variable oder eine Kombination aus Variablen enthalten, die jeden Fall (Kunden) identifizieren.

Erstellen von RFM-Scores aus Kundendaten

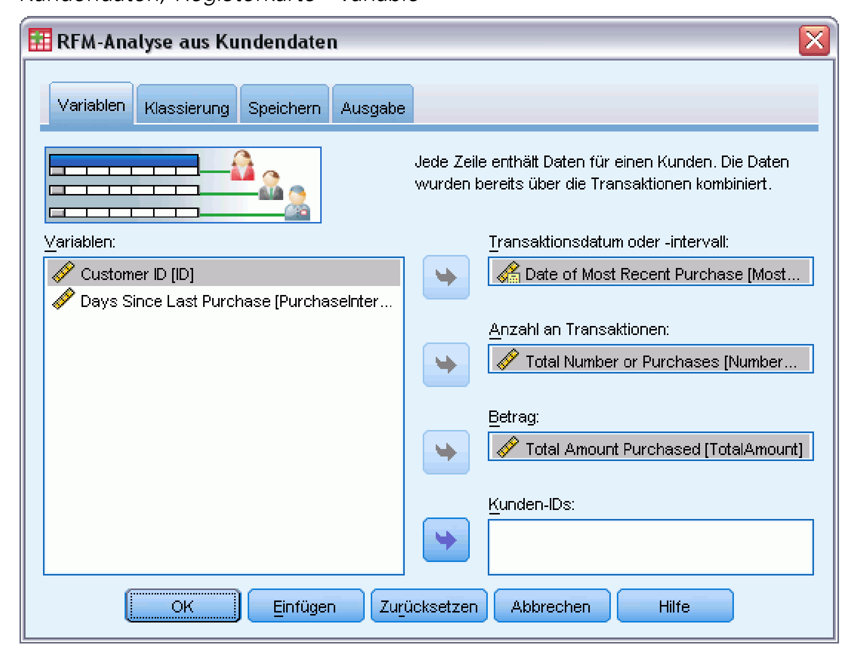
▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Hilfe bei der Erkennung meiner besten Kontakte (RFM-Analyse) und klicken Sie auf Weiter.

▶ Wählen Sie Kundendaten und klicken Sie auf Weiter.

Abbildung 2-5
Kundendaten, Registerkarte "Variable"



- Wählen Sie die Variable, die das letzte Transaktionsdatum oder eine Zahl enthält, die ein Zeitintervall seit der letzten Transaktion darstellt.

- Wählen Sie die Variable, die die Gesamtzahl der Transaktionen für jeden Kunden enthält.

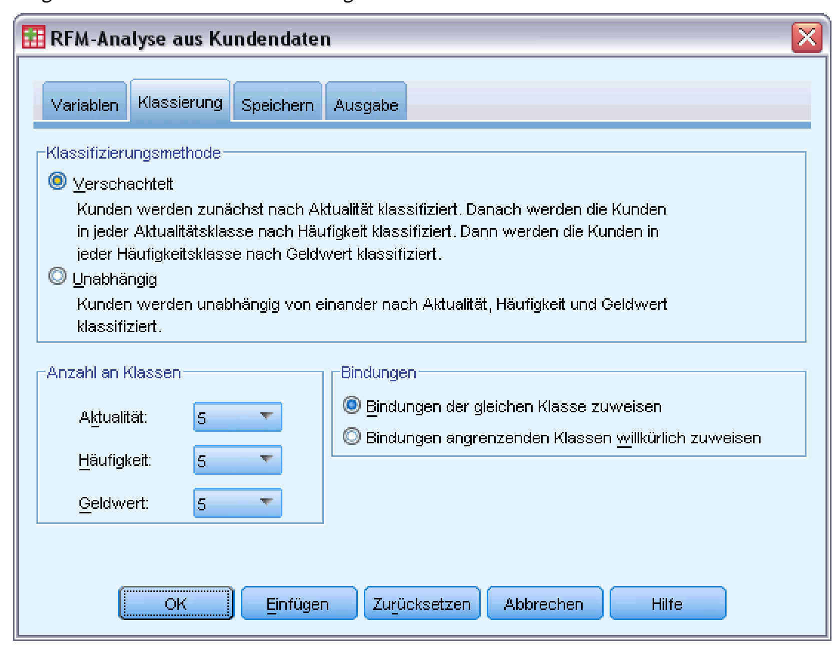
- Wählen Sie die Variable, die den Gesamtgeldwertbetrag für jeden Kunden enthält.

- Wenn Sie die RFM-Scores in ein neues Daten-Set schreiben möchten, wählen Sie die Variable oder die Kombination aus Variablen, die jeden Kunden eindeutig identifiziert. Zum Beispiel könnten Fälle durch einen eindeutigen Schlüsselcode oder eine Kombination aus Nachname und Vorname identifiziert werden.

RFM-Klassifizierung

Der Prozess der Gruppierung einer großen Zahl von numerischen Werten in eine kleine Zahl von Kategorien wird manchmal als **Klassifizierung** (Binning) bezeichnet. Bei der RFM-Analyse sind die Klassen Einstufungskategorien. Sie können die Registerkarte "Klassifizierung" verwenden, um die zur Zuweisung von Aktualitäts-, Häufigkeits- und Geldwertwerten zu diesen Klassen verwendete Methode zu ändern.

Abbildung 2-6
Registerkarte "RFM-Klassierung"



Klassifizierungsmethode

Verschachtelt. Bei der verschachtelten Klassifizierung wird den Aktualitätswerten eine einfache Einstufung zugewiesen. Innerhalb jeder Aktualitätseinstufung wird Kunden eine Häufigkeitseinstufung zugewiesen. Innerhalb jeder Häufigkeitseinstufung wird Kunden eine Geldwerteinstufung zugewiesen. Diese neigt dazu, eine gleichmäßigere Verteilung von kombinierten RFM-Scores bereitzustellen, hat jedoch den Nachteil, dass sich die Interpretation der Häufigkeits- und Geldwerteinstufungs-Scores schwieriger gestaltet. Zum Beispiel kann eine Häufigkeitseinstufung von 5 für einen Kunden mit einer Aktualitätseinstufung von 5 nicht das Gleiche bedeuten wie eine Häufigkeitseinstufung von 5 für einen Kunden mit einer Aktualitätseinstufung von 4, denn die Häufigkeitseinstufung hängt von der Aktualitätseinstufung ab.

Unabhängig. Aktualitäts-, Häufigkeits- und Geldwerte werden einfachen Einstufungen zugewiesen. Die drei Einstufungen werden unabhängig zugewiesen. Die Interpretation jeder der drei RFM-Komponenten ist daher eindeutig. Ein Häufigkeits-Score von 5 für einen Kunden bedeutet das Gleiche wie ein Häufigkeits-Score von 5 für einen anderen Kunden, unabhängig von ihren Aktualitäts-Scores. Bei kleineren Stichproben hat dies den Nachteil, dass es zu einer weniger gleichmäßigen Verteilung der kombinierten RFM-Scores kommt.

Anzahl an Klassen

Die Anzahl der Kategorien (Klassen) für jede Komponente für die Erstellung der RFM-Scores. Die Gesamtzahl der möglichen kombinierten RFM-Scores ist das Produkt der drei Werte. Zum Beispiel würden 5 Aktualitätsklassen, 4 Häufigkeitsklassen und 3 Geldwertklassen insgesamt 60 mögliche kombinierte RFM-Scores zwischen 111 und 543 erzeugen.

- Standard für jede Komponente ist 5, so dass 125 mögliche, kombinierte RFM-Scores zwischen 111 und 555 erzeugt werden.
- Die maximale Zahl an zulässigen Klassen für jede Score-Komponente ist neun.

Bindungen

Eine "Bindung" sind einfach zwei oder mehr gleiche Aktualitäts-, Häufigkeits- oder Geldwerte. Idealerweise wünscht man sich ungefähr die gleiche Zahl an Kunden in jeder Klasse, aber eine größere Zahl an Bindungswerten kann sich auf die Klassenverteilung auswirken. Es gibt zwei Alternativen für die Handhabung von Bindungen:

- **Bindungen der gleichen Klasse zuweisen**. Diese Methode weist unabhängig von der Auswirkung auf die Klassenverteilung gebundene Werte stets der gleichen Klasse zu. So ergibt sich eine konsistente Klassifizierungsmethode: Wenn zwei Kunden den gleichen Aktualitätswert besitzen, werden sie stets dem gleichen Aktualitäts-Score zugewiesen. In einem extremen Beispiel haben Sie vielleicht 1.000 Kunden, von denen 500 ihren letzten Einkauf am gleichen Tag tätigen. In einer 5-Klassen-Einstufung würden 50 % der Kunden daher anstelle der gewünschten 20 % einen Aktualitäts-Score von 5 erhalten.

  Beachten Sie, dass es bei der verschachtelten Klassifizierungsmethode "Konsistenz" bei Häufigkeits- und Geldwert-Scores etwas komplizierter ist, da Häufigkeits-Scores innerhalb von Aktualitäts-Score-Klassen und Geldwert-Scores innerhalb von Häufigkeits-Score-Klassen zugewiesen werden. So haben zwei Kunden mit dem gleichen Häufigkeitswert ggf. nicht den gleichen Häufigkeits-Score, wenn sie nicht, unabhängig von der Handhabung gebundener Werte, auch über den gleichen Aktualitäts-Score verfügen.
- **Bindungen willkürlich zuweisen.** Hierüber wird eine gleichmäßige Klassenverteilung gewährleistet, indem Bindungen vor der Einstufung ein sehr kleiner Varianzfaktor zugewiesen wird, so dass es zum Zweck der Zuweisung von Werten an die eingestuften Klassen keine gebundenen Werte gibt. Dieser Prozess hat keine Auswirkungen auf die Originalwerte. Er wird nur eingesetzt, um Bindungen eindeutig zu machen. Zwar erzeugt dies eine gleichmäßige Klassenverteilung (ungefähr die gleiche Anzahl an Kunden in jeder Klasse), es kann aber auch zu vollständig unterschiedlichen Score-Ergebnissen für Kunden führen, die ähnliche oder identische Aktualitäts-, Häufigkeits- oder Geldwerte haben, speziell, wenn die Anzahl der Kunden relativ klein und/oder die Anzahl der Bindungen relativ hoch ist.

Tabelle 2-1
Bindungen der gleichen Klasse zuweisen im Vergleich mit Bindungen willkürlich zuweisen

ID	Letzter Kauf (Aktualität)	Aktualitätseinstufung	
		Bindungen der gleichen Klasse zuweisen	Bindungen willkürlich zuweisen
1	29.10.2006	5	5
2	28.10.2006	4	4
3	28.10.2006	4	4
4	28.10.2006	4	5
5	28.10.2006	4	3
6	21.9.2006	3	3

ID	Letzter Kauf (Aktualität)	Aktualitätseinstufung	
		Bindungen der gleichen Klasse zuweisen	Bindungen willkürlich zuweisen
7	21.9.2006	3	2
8	13.8.2006	2	2
9	13.8.2006	2	1
10	20.6.2006	1	1

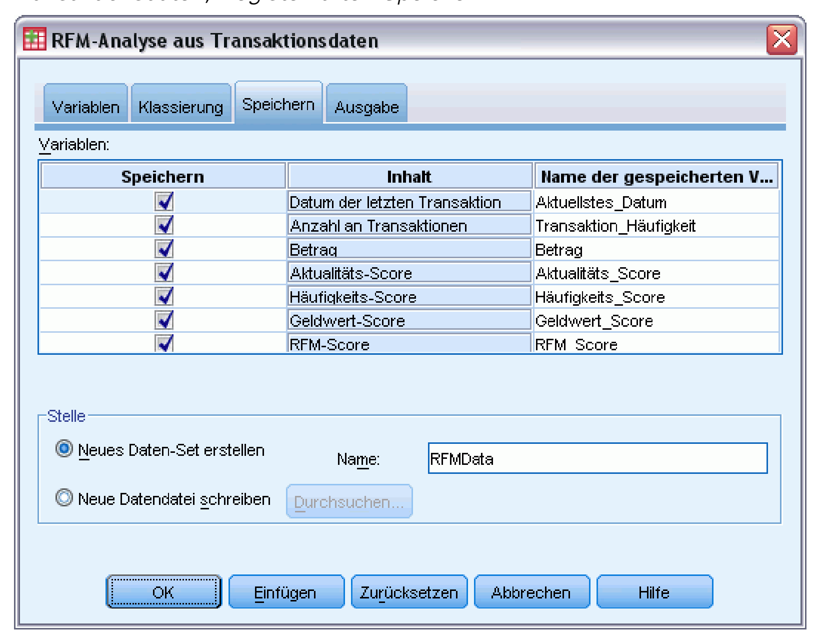
- In diesem Beispiel führt die Zuweisung von Bindungen der gleichen Klasse zu einer ungleichmäßigen Klassenverteilung: 5 (10 %), 4 (40 %), 3 (20 %), 2 (20 %), 1 (10 %).
- Das willkürlich Zuweisen von Bindungen führt zu 20 % in jeder Klasse. Um dieses Ergebnis aber zu erreichen, werden die vier Fälle mit dem Datumswert 28.10.2006 3 verschiedenen Klassen zugewiesen und die 2 Fälle mit einem Datumswert von 13.8.2006 werden ebenfalls unterschiedlichen Klassen zugewiesen.

  Beachten Sie, dass die Art, mit der Bindungen unterschiedlichen Klassen zugewiesen werden, absolut zufällig erfolgt innerhalb der Einschränkung, dass das Endergebnis eine gleiche Anzahl von Fällen in jeder Klasse hat). Wenn Sie eine zweite Menge an Scores mit der gleichen Methode berechnet haben, könnte sich die Einstufung für einen bestimmten Fall mit einem gebundenen Wert ändern. Zum Beispiel könnten sich die Einstufungen von 5 und 3 für die Fälle 4 und 5 beim zweiten Mal vertauschen.

Speichern von RFM-Scores aus Transaktionsdaten

RFM aus Transaktionsdaten erstellt stets ein neues aggregiertes Daten-Set mit einer Zeile je Kunde. Verwenden Sie die Registerkarte “Speichern”, um anzugeben, welche Scores und anderen Variablen Sie speichern möchten und wo Sie sie speichern möchten.

Abbildung 2-7
Transaktionsdaten, Registerkarte "Speichern"



Variablen

Die Schlüsselvariablen, die jeden Kunden eindeutig identifizieren, werden automatisch im neuen Daten-Set gespeichert. Die folgenden zusätzlichen Variablen können im neuen Daten-Set gespeichert werden:

- **Datum der letzten Transaktion für jeden Kunden.**
- **Anzahl der Transaktionen.** Die Gesamtzahl an Transaktionszeilen je Kunde.
- **Betrag.** Der Gesamtbetrag für jeden Kunden, basierend auf der in der Registerkarte "Variablen" gewählten Auswertungsmethode.
- **Aktualitäts-Score.** Der jedem Kunden zugewiesene Score, basierend auf dem letzten Transaktionsdatum. Höhere Scores geben aktuellere Transaktionsdaten an.
- **Häufigkeits-Score.** Der jedem Kunden zugewiesene Score, basierend auf der Gesamtzahl an Transaktionen. Höhere Scores stehen für mehr Transaktionen.
- **Geldwert-Score.** Der jedem Kunden zugewiesene Score, basierend auf der ausgewählten Geldwert-Auswertungskennzahl. Höhere Scores stehen für einen höheren Wert für die Geldwert-Auswertungskennzahl.
- **RFM-Score.** Die drei Einzel-Scores, zu einem einzigen Wert kombiniert: *(Aktualität x 100) + (Häufigkeit x 10) + Geldwert*.

Standardmäßig werden alle verfügbaren Variablen in das neue Daten-Set aufgenommen. Deaktivieren Sie die, die Sie nicht aufnehmen möchten. Optional können Sie Ihre eigenen Variablennamen angeben. Die Variablennamen müssen den Regeln zum Benennen von Variablen entsprechen.

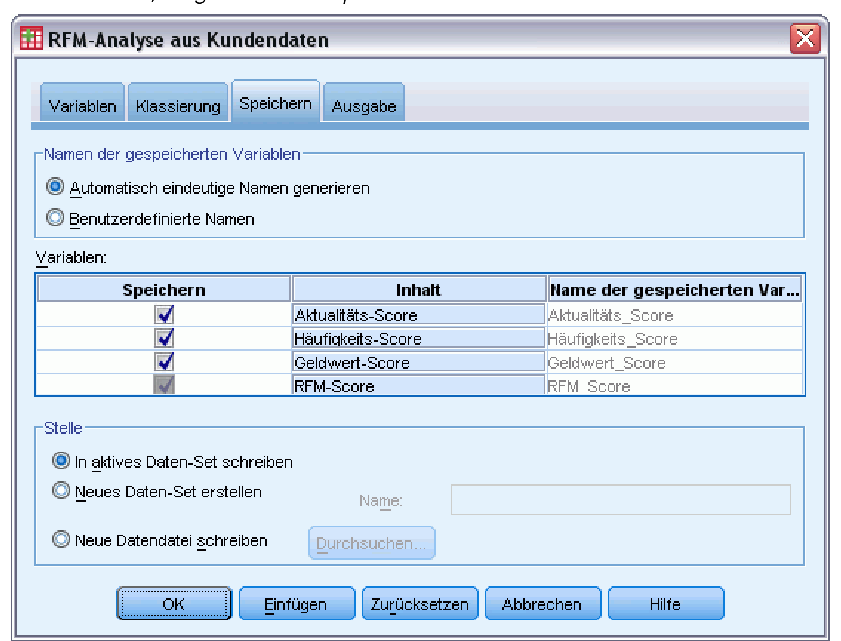
Ort

RFM aus Transaktionsdaten erstellt stets ein neues aggregiertes Daten-Set mit einer Zeile je Kunde. Sie können ein neues Daten-Set in der aktuellen Sitzung erstellen oder die RFM-Score-Daten in einer externen Datendatei speichern. Die Namen von Daten-Sets müssen den Regeln zum Benennen von Variablen entsprechen. (Diese Beschränkung gilt nicht für Namen von externen Datendateien.)

Speichern von RFM-Scores aus Kundendaten

Für Kundendaten können Sie die RFM-Score-Variablen dem aktiven Daten-Set hinzufügen oder ein neues Daten-Set erstellen, das die ausgewählten Score-Variablen enthält. Verwenden Sie die Registerkarte "Speichern", um anzugeben, welche Score-Variablen Sie speichern möchten und wo Sie sie speichern möchten.

Abbildung 2-8
Kundendaten, Registerkarte "Speichern"



Name der gespeicherten Variablen

- **Generieren Sie automatisch eindeutige Namen.** Wenn Sie Score-Variablen dem aktiven Daten-Set hinzufügen, stellt diese Option sicher, dass neue Variablennamen eindeutig sind. Dies ist besonders nützlich, wenn Sie dem aktiven Daten-Set mehrere unterschiedliche Sets an RFM-Scores (basierend auf unterschiedlichen Kriterien) hinzufügen möchten.
- **Benutzerdefinierte Namen.** Über diese Option können Sie den Score-Variablen Ihre eigenen Variablennamen zuweisen. Die Variablennamen müssen den Regeln zum Benennen von Variablen entsprechen.

Variablen

Wählen (aktivieren) Sie die Variablen, die Sie speichern möchten:

- **Aktualitäts-Score.** Der jedem Kunden zugewiesene Score, basierend auf dem Wert des Transaktionsdatums oder der Intervallvariablen, die in der Registerkarte “Variablen” ausgewählt ist. Höhere Scores werden neueren Daten bzw. niedrigeren Intervallwerten zugewiesen.
- **Häufigkeits-Score.** Der jedem Kunden zugewiesene Score, basierend auf der Variablen “Anzahl der Transaktionen”, die in der Registerkarte “Variablen” ausgewählt ist. Höhere Scores werden höheren Werten zugewiesen.
- **Geldwert-Score.** Der jedem Kunden zugewiesene Score, basierend auf der Variablen “Betrag”, die in der Registerkarte “Variablen” ausgewählt ist. Höhere Scores werden höheren Werten zugewiesen.
- **RFM-Score.** Die drei Einzel-Scores, zu einem einzigen Wert kombiniert: *(Aktualität*100)+(Häufigkeit*10)+Geldwert.*

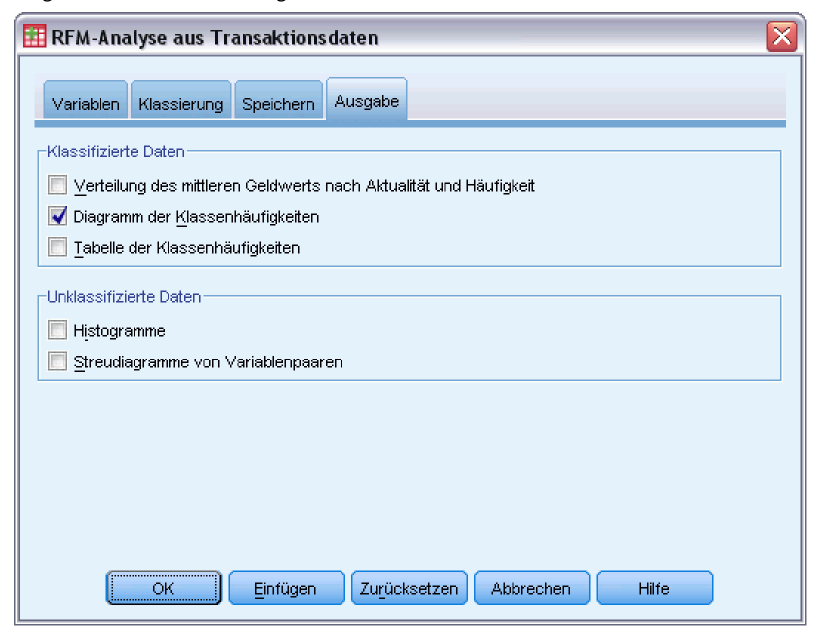
Ort

Für Kundendaten gibt es drei Alternativen für den Speicherort von neuen RFM-Scores:

- **Aktives Daten-Set.** Ausgewählte RFM-Score-Variablen werden dem aktiven Daten-Set hinzugefügt.
- **Neues Daten-Set.** Ausgewählte RFM-Score-Variablen und die Schlüsselvariablen, die jeden Kunden (Fall) eindeutig identifizieren, werden in ein neues Daten-Set in der aktuellen Sitzung geschrieben. Die Namen von Daten-Sets müssen den Regeln zum Benennen von Variablen entsprechen. Diese Option ist nur verfügbar, wenn Sie eine oder mehrere Kunden-ID-Variablen in der Registerkarte “Variablen” wählen.
- **Datei.** Ausgewählte RFM-Scores und die Schlüsselvariablen, die jeden Kunden (Fall) eindeutig identifizieren, werden in einer externen Datendatei gespeichert. Diese Option ist nur verfügbar, wenn Sie eine oder mehrere Kunden-ID-Variablen in der Registerkarte “Variablen” wählen.

RFM-Ausgabe

Abbildung 2-9
Registerkarte "RFM-Ausgabe"



Klassifizierte Daten

Diagramme und Tabellen für klassifizierte Daten basieren auf den berechneten Aktualitäts-, Häufigkeits- und Geldwert-Scores.

Verteilung des mittleren Geldwerts nach Aktualität und Häufigkeit. Die Verteilung des mittleren Geldwerts zeigt den durchschnittlichen Geldwert für Kategorien, die durch Aktualitäts- und Häufigkeits-Scores definiert sind. Dunklere Bereiche zeigen einen höheren durchschnittlichen Geldwert an.

Diagramm der Klassenhäufigkeiten. Das Diagramm der Klassenhäufigkeiten zeigt die Klassenverteilung für die ausgewählten Klassifizierungsmethoden an. Jeder Balken steht für die Anzahl der Fälle, die jedem kombinierten RFM-Score zugewiesen werden.

- Auch wenn Sie sich in der Regel eine relativ gleichmäßige Verteilung wünschen, bei der alle (oder die meisten) Balken ungefähr die gleiche Höhe haben, sollte eine gewisse Varianz erwartet werden, wenn die Standard-Klassifizierungsmethode verwendet wird, die gebundene Werte der gleichen Klasse zuweist.
- Extreme Schwankungen in der Klassenverteilung und/oder viele leere Klassen können anzeigen, dass Sie eine andere Klassifizierungsmethode (weniger Klassen und/oder zufällige Zuweisung von Bindungen) versuchen oder die Eignung der RFM-Analyse überdenken sollten.

Tabelle der Klassenhäufigkeiten. Die gleichen Informationen, die sich im Diagramm der Klassenhäufigkeiten finden, nur in Form einer Tabelle mit Klassenhäufigkeiten in jeder Zelle.

Unklassifizierte Daten

Diagramme und Tabellen für unklassifizierte Daten basieren auf den Originalvariablen, die für die Erstellung der Aktualitäts-, Häufigkeits- und Geldwert-Scores verwendet wurden.

Histogramme. Die Histogramme zeigen die relative Verteilung von Werten für die drei Variablen, die für die Berechnung der Aktualitäts-, Häufigkeits- und Geldwert-Scores verwendet wurden. Diese Histogramme zeigen oftmals etwas verzerrte Verteilungen anstelle einer normalen oder symmetrischen Verteilung an.

Die horizontale Achse jedes Histogramms ist stets von niedrigeren Werten links zu hohen Werten rechts geordnet. Bei der Aktualität hängt jedoch die Interpretation des Diagramms vom Typ der Aktualitätsmessung ab: Datum und Zeitintervall. Für Daten stellen die Balken links Werte dar, die weiter in der Vergangenheit liegen (ein weniger aktuelles Datum hat einen geringeren Wert als ein aktuelleres Datum). Für Zeitintervalle stellen die Balken links aktuellere Werte dar (je kleiner das Zeitintervall, umso aktueller die Transaktion).

Streudiagramme von Variablenpaaren. Diese Streudiagramme zeigen die Beziehungen zwischen den drei Variablen, die für die Berechnung der Aktualitäts-, Häufigkeits- und Geldwert-Scores verwendet wurden.

Es ist bei Streudiagrammen für Häufigkeit und Geldwerte üblich, eine positive Korrelation zu zeigen, da der Geldwert den Gesamtbetrag aller Transaktionen darstellt. Eine größere Zahl an Transaktionen führt wahrscheinlich zu einem größeren Gesamtbetrag. Es ist zudem üblich, eine wahrnehmbare lineare Gruppierung der Punkte auf der Häufigkeitsskala festzustellen, da die Häufigkeit oftmals einen relativ kleinen Bereich diskreter Werte darstellt. Wenn zum Beispiel die Gesamtzahl der Transaktionen 15 nicht überschreitet, gibt es nur 15 mögliche Häufigkeitswerte (außer Sie zählen teilweise Transaktionen), während es Hunderte von möglichen Aktualitäts- und Tausende von Geldwerten geben kann.

Die Interpretation der Aktualitätsachsen hängt jedoch vom Typ der Aktualitätsmessung ab: Datum und Zeitintervall. Bei Daten stellen Punkte näher am Ursprung Daten dar, die weiter in der Vergangenheit liegen. Bei Zeitintervallen stellen Punkte näher am Ursprung aktuellere Werte dar.

Kapitel

3

Cluster-Analyse

Bei der Cluster-Analyse handelt es sich um eine explorative Prozedur zum Ermitteln von natürlichen Gruppierungen (Clustern) innerhalb Ihrer Daten. Damit können beispielsweise verschiedene Kundengruppen auf der Basis unterschiedlicher demographischer und Kaufverhaltensmerkmale ausgemacht werden.

Beispiel. In Einzel- und Fachhandel werden Cluster-Methoden regelmäßig auf Daten angewendet, die Kaufgewohnheiten, Geschlecht, Alter und Einkommensniveau der Kundschaft beschreiben. Ziel der Analyse ist eine Ausrichtung der unternehmenseigenen Marketing- und Produktentwicklungsstrategien auf einzelne Konsumentengruppen, um Umsatzsteigerungen und Markentreue zu erreichen.

Erläuterungen der Daten für die Clusteranalyse

Daten. Mit dieser Prozedur können sowohl stetige als auch kategoriale Felder analysiert werden. Jeder Datensatz (Zeile) stellt einen Kunden dar, der gruppiert werden soll, während die Felder (Variablen) die Attribute darstellen, auf deren Grundlage die Gruppierung erfolgt.

Datensatz-Reihenfolge. Beachten Sie, dass die Ergebnisse von der Reihenfolge der Datensätze abhängen können. Um die Auswirkungen der Reihenfolge zu minimieren, sollten Sie versuchen, die Datensätze in zufälliger Reihenfolge zu mischen. Prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie die Analyse mehrmals durchführen, wobei die Datensätze in einer unterschiedlichen, zufällig ausgewählten Reihenfolge sortiert sind.

Messniveau. Es ist wichtig, das korrekte Messniveau zuzuweisen, da sich dieses auf die Berechnung der Ergebnisse auswirkt.

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Kontinuierlich.** Eine Variable kann als metrisch (stetig) behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

Ein Symbol neben jedem Feld zeigt das aktuelle Messniveau an.

Messniveau	**Datentyp**			
	Numerisch	Zeichenfolge	Datum	Zeit
Metrisch (stetig)		entfällt		
Ordinal				
Nominal				

Sie können das Messniveau in der Variablenansicht des Daten-Editors ändern oder das Dialogfeld “Variableneigenschaften definieren” verwenden, um ein geeignetes Messniveau für jedes Feld anzugeben .

So führen Sie eine Clusteranalyse durch

Wählen Sie die folgenden Befehle aus den Menüs aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Meine Kontakte in Cluster segmentieren aus.

Abbildung 3-1
Registerkarte "Clusteranalysefelder"

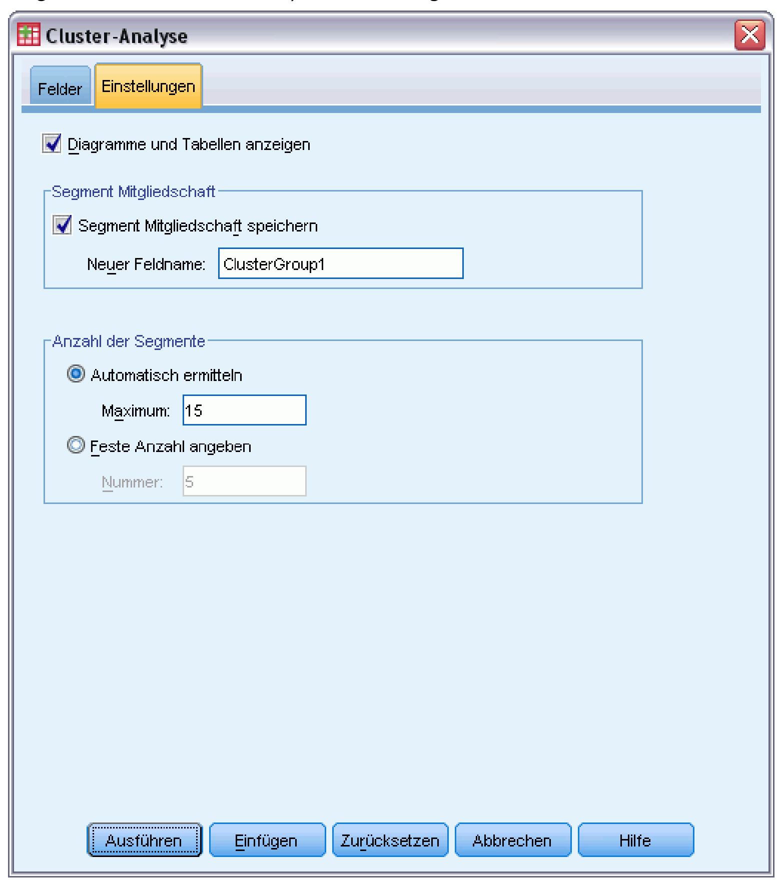


- Wählen Sie die kategorialen (nominalen, ordinalen) und stetigen (metrischen) Felder aus, die Sie zum Erstellen von Segmenten verwenden möchten.

- Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Einstellungen

Abbildung 3-2
Registerkarte "Clusteranalyseeinstellungen"



Auf der Registerkarte "Einstellungen" können Sie Diagramme und Tabellen, die die Segmente beschreiben, anzeigen oder unterdrücken, ein neues Feld im Daten-Set speichern, das das Segment (Cluster) für jeden Datensatz im Daten-Set identifiziert, und festlegen, wie viele Segmente die Cluster-Lösung enthalten soll.

Diagramme und Tabellen anzeigen. Zeigt Tabellen und Diagramme an, die die Segmente beschreiben.

Segment-Zugehörigkeit. Speichert ein neues Feld bzw. eine neue Variable, das bzw. die das Segment identifiziert, zu dem jeder Datensatz gehört.

- Die Feldnamen müssen den Benennungsregeln von PASW Statistics entsprechen.
- Der Feldname der Segment-Zugehörigkeit kann kein Duplikat eines Feldnamens sein, der bereits im Daten-Set vorhanden ist. Falls Sie diese Prozedur also mehr als einmal mit demselben Daten-Set ausführen, müssen Sie jedes Mal einen anderen Namen angeben.

- **Anzahl der Segmente.** Legt fest, wie die Anzahl der Segmente ermittelt wird.
- **Automatisch ermitteln.** Die Prozedur ermittelt automatisch die “beste” Anzahl der Segmente bis zum angegebenen Höchstwert.

Feste Anzahl angeben. Die Prozedur erzeugt die angegebene Anzahl der Segmente.

Kapitel



Profile über potenzielle Kunden

Bei dieser Technik werden Ergebnisse aus einer früheren Kampagne oder einer Testkampagne verwendet, um beschreibende Profile zu erstellen. Diese Profile können bei zukünftigen Kampagnen für das Targeting bestimmter Gruppen von Kontakten verwendet werden. Das Responsefeld zeigt, wer auf die frühere Kampagne bzw. die Testkampagne reagiert hat. Die Liste "Profile" enthält die Merkmale, die Sie zur Erstellung des Profils verwenden möchten.

Beispiel. Anhand der Ergebnisse einer Testsendung möchte die Marketing-Abteilung eines Unternehmens auf Basis von demographischen Informationen Profile der Typen von Kunden erstellen, bei denen die Wahrscheinlichkeit einer Antwort auf ein Angebot am höchsten ist.

Ausgabe

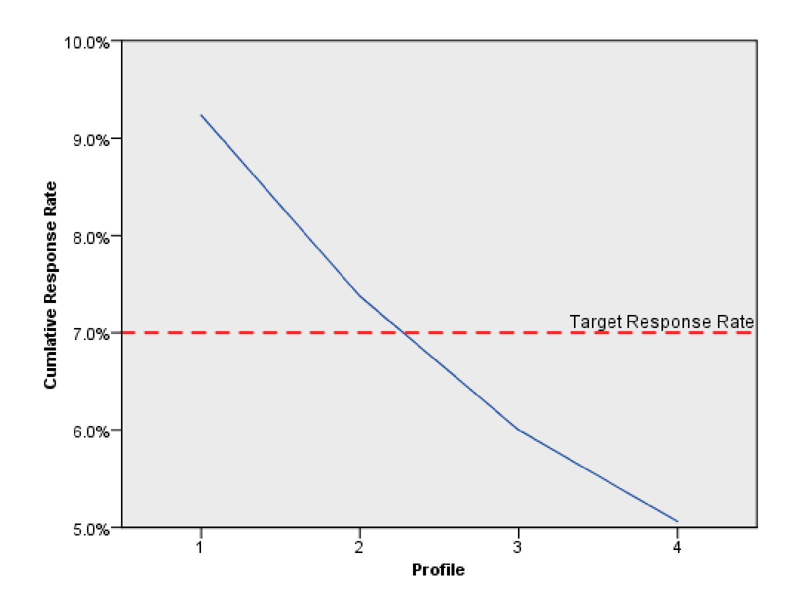
Die Ausgabe enthält eine Tabelle, die eine Beschreibung jeder Profilgruppe enthält und in der Responseraten (Prozentsatz der positiven Antworten), kumulative Responseraten sowie ein Diagramm der kumulativen Responseraten angezeigt werden. Wenn Sie eine minimale Zielresponserate einschließen, wird die Tabelle farbkodiert, so dass erkennbar ist, welche Profile der Mindestanforderung an die kumulative Responserate entsprechen. Das Diagramm enthält eine Bezugslinie, die den Wert der minimalen Responserate kenntlich macht.

Abbildung 4-1
Tabelle und Diagramm für die Responserate

Responserate

Nummer	Profil			
	Beschreibung	Gruppengröße	Responserate	Kumulierte Responserate
1	Region = "West","South","East" Gender = "Female" Married = "No"	379	9.2%	9.2%
2	Region = "West","South","East" Gender = "Female" Married = "Yes"	299	5.0%	7.4%
3	Region = "West","South","East" Gender = "Male"	722	4.7%	6.0%
4	Region = "North"	517	2.5%	5.1%

Grün: Erfüllt die Ziel-Responserate.
Rot: Erfüllt die Ziel-Responserate nicht.



Erläuterung der Daten für Profile über potenzielle Kunden

Responsefeld. Das Responsefeld muss nominal oder ordinal sein. Es kann ein numerisches Feld oder ein String-Feld sein. Falls dieses Feld einen Wert enthält, der die Anzahl von Käufen anzeigt, müssen Sie ein neues Feld erstellen, in dem ein einzelner Wert sämtliche positiven Antworten repräsentiert. Für weitere Informationen siehe Erstellen eines kategorialen Responsefelds auf S. 24.

Wert für positive Antworten. Der Wert für positive Antworten bezeichnet diejenigen Kunden, die positiv reagiert haben (zum Beispiel, indem sie einen Kauf getätigt haben). Es wird davon ausgegangen, dass alle anderen nicht fehlenden Responsewerte eine negative Antwort anzeigen. Falls es definierte Wertelabels für das Responsefeld gibt, werden diese Labels in der Dropdown-Liste angezeigt.

Profile erstellen mit. Diese Felder können nominal, ordinal oder stetig (metrisch) sein. Es können numerische Felder oder String-Felder sein.

Messniveau. Es ist wichtig, das korrekte Messniveau zuzuweisen, da sich dieses auf die Berechnung der Ergebnisse auswirkt.

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Kontinuierlich.** Eine Variable kann als metrisch (stetig) behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

Ein Symbol neben jedem Feld zeigt das aktuelle Messniveau an.

Messniveau	**Datentyp**			
	Numerisch	Zeichenfolge	Datum	Zeit
Metrisch (stetig)		entfällt		
Ordinal				
Nominal				

Sie können das Messniveau in der Variablenansicht des Daten-Editors ändern oder das Dialogfeld “Variableneigenschaften definieren” verwenden, um ein geeignetes Messniveau für jedes Feld anzugeben .

So erhalten Sie Profile über potenzielle Kunden

Wählen Sie die folgenden Befehle aus den Menüs aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Profile für die Kontakte erstellen, die auf ein Angebot reagiert haben.

Abbildung 4-2
Profile über potenzielle Kunden, Registerkarte "Felder"



▶ Wählen Sie das Feld aus, das anzeigt, welche Kontakte auf das Angebot reagiert haben. Dieses Feld muss nominal oder ordinal sein.

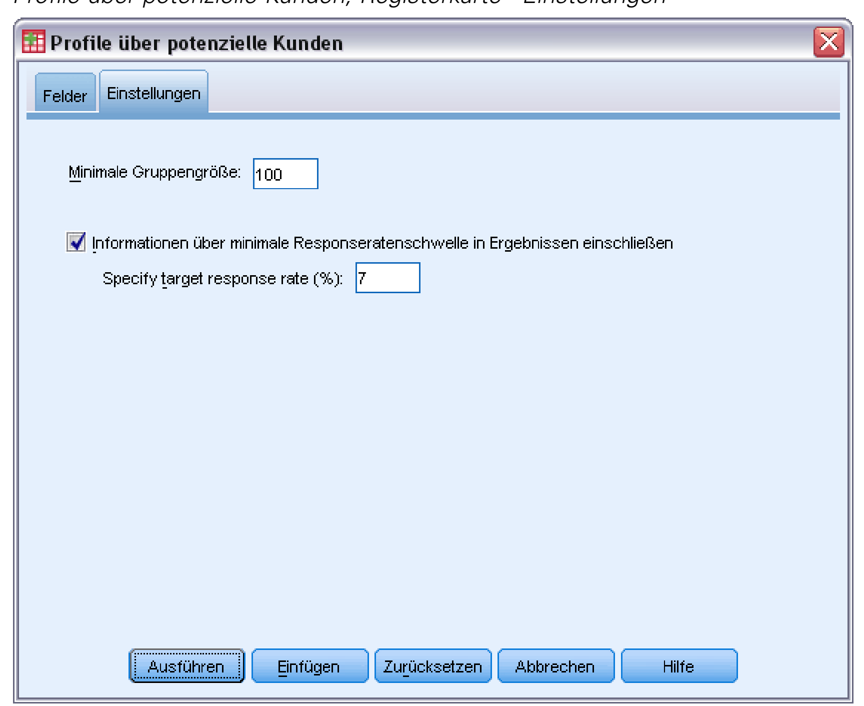
▶ Geben Sie den Wert an, der eine positive Antwort anzeigt. Für Werte mit definierten Wertelabels können Sie das Wertelabel aus der Dropdown-Liste auswählen, woraufhin der entsprechende Wert angezeigt wird.

▶ Wählen Sie die Felder aus, die Sie verwenden möchten, um die Profile zu erstellen.

▶ Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Einstellungen

Abbildung 4-3
Profile über potenzielle Kunden, Registerkarte "Einstellungen"



Auf der Registerkarte "Einstellungen" können Sie die Mindestgröße der Profilgruppe angeben und eine minimale Responseratenschwelle in die Ausgabe einschließen.

Minimale Profilgruppengröße. Jedes Profil repräsentiert die gemeinsamen Merkmale einer Gruppe von Kontakten im Daten-Set (z.B. Frauen unter 40 Jahren, die im Westen leben). Standardmäßig ist 100 der kleinste Wert für die Größe der Profilgruppe. Kleinere Gruppengrößen können zu einer größeren Anzahl von Gruppen führen, größere Gruppengrößen liefern jedoch verlässlichere Ergebnisse. Dieser Wert muss eine positive Ganzzahl sein.

Informationen über minimale Responseratenschwelle in Ergebnissen einschließen. Die Ergebnisse enthalten eine Tabelle, in der Responseraten (Prozentsatz der positiven Antworten), kumulative Responseraten sowie ein Diagramm der kumulativen Responseraten angezeigt werden. Wenn Sie eine minimale Zielresponserate eingeben, wird die Tabelle farbkodiert, so dass erkennbar ist, welche Profile der Mindestanforderung an die kumulative Responserate entsprechen. Das Diagramm enthält eine Bezugslinie, die den Wert der minimalen Responserate kenntlich macht. Der Wert muss größer als 0 und kleiner als 100 sein.

Erstellen eines kategorialen Responsefelds

Das Responsefeld sollte kategorial sein, wobei ein Wert alle positiven Reaktionen darstellen sollte. Es wird davon ausgegangen, dass alle anderen nicht fehlenden Werte eine negative Antwort anzeigen. Falls das Responsefeld einen stetigen (metrischen) Wert enthält, beispielsweise die

Anzahl oder den Geldwert der Käufe, müssen Sie ein neues Feld erstellen, das allen von Null abweichenden Responsewerten eine einzelne positive Antwort zuweist.

- Falls negative Antworten als "0" (nicht leer, was als fehlender Wert aufgefasst wird) aufgezeichnet werden, kann dies anhand der folgenden Formel berechnet werden:

  NewName=OldName>0,

  wobei *NewName* der Name des neuen Felds und *OldName* der Name des neuen Felds ist. Dies ist ein logischer Ausdruck, der allen nicht fehlenden Werten größer 0 einen Wert von 1 und allen nicht fehlenden Werten kleiner oder gleich 0 den Wert 0 zuweist.
- Falls für negative Antworten kein Wert aufgezeichnet wird, werden diese Werte als fehlend behandelt und die Formel ist etwas komplizierter:

  NewName=NOT(MISSING(OldName))

  Bei diesem logischen Ausdruck wird allen nicht fehlenden Responsewerten ein Wert von 1 und allen fehlenden Responsewerten ein Wert von 0 zugewiesen.
- Falls Sie zwischen negativen (0) Responsewerten und fehlenden Werten nicht unterscheiden können, kann kein korrekter Responsewert berechnet werden. Falls es nur relativ wenig tatsächlich fehlende Werte gibt, muss dies jedoch keine großen Auswirkungen auf die berechneten Responseraten haben. Falls es jedoch viele fehlende Werte gibt – z. B. wenn die Responseinformationen nur für eine kleine Teststichprobe des gesamten Daten-Sets berechnet werden –, wird dies dazu führen, dass die berechneten Responseraten bedeutungslos sind, da sie deutlich niedriger sein werden als die tatsächlichen Responseraten.

So erstellen Sie ein kategoriales Responsefeld

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Transformieren
  Variable berechnen

▶ Geben Sie für "Zielvariable" einen neuen Feld-(Variablen-)Namen ein.

▶ Falls negative Reaktionen als 0 aufgezeichnet werden, geben Sie als numerischen Ausdruck OldName>0 ein, wobei *OldName* der ursprüngliche Feldname ist.

▶ Falls negative Reaktionen als fehlend (leer) aufgezeichnet werden, geben Sie als numerischen Ausdruck NOT(MISSING(OldName)) ein, wobei *OldName* der ursprüngliche Feldname ist.

Kapitel



Responseraten nach Postleitzahlen

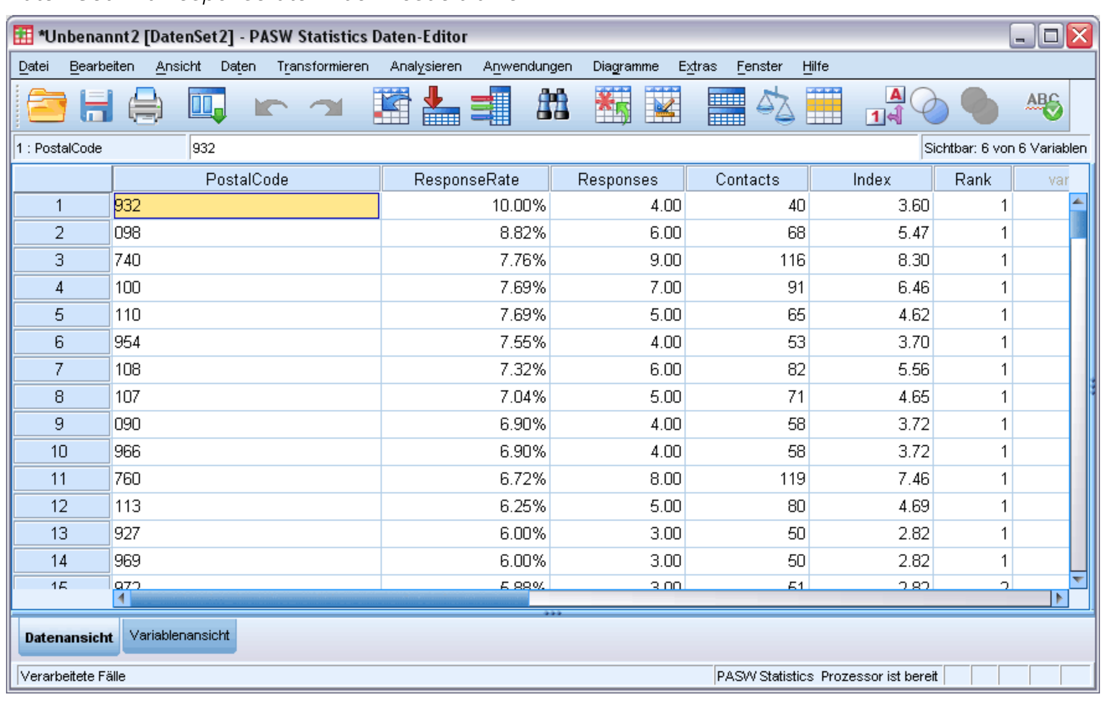
Bei dieser Technik werden Ergebnisse aus einer früheren Kampagne verwendet, um Responseraten nach Postleitzahlen zu berechnen. Diese Raten können bei zukünftigen Kampagnen für das Targeting bestimmter Postleitzahlbereiche verwendet werden. Das Responsefeld zeigt an, wer auf die frühere Kampagne reagiert hat. Das Postleitzahlfeld kennzeichnet das Feld, das die Postleitzahlen enthält.

Beispiel. Anhand der Ergebnisse einer früheren Postsendungs-Kampagne erzeugt die Marketing-Abteilung eines Unternehmens Responseraten nach Postleitzahlen. Auf Basis verschiedener Kriterien wie der minimalen akzeptablen Responserate und/oder der maximalen Anzahl von Kontakten, die in die Postsendungs-Kampagne eingeschlossen werden sollen, können daraufhin bestimmte Postleitzahlbereiche für die Kampagne bestimmt werden.

Ausgabe

Zur Ausgabe dieser Prozedur gehört ein neues Daten-Set, das die Responseraten nach Postleitzahl sowie eine Tabelle und ein Diagramm enthält, die die Ergebnisse nach Dezil-Rang zusammenfassen (oberste 10 %, oberste 20 % usw.). Die Tabelle kann auf Basis einer vom Benutzer festgelegten minimalen kumulativen Responserate oder maximalen Anzahl von Kontakten farbkodiert werden.

Abbildung 5-1
Daten-Set mit Responseraten nach Postleitzahlen



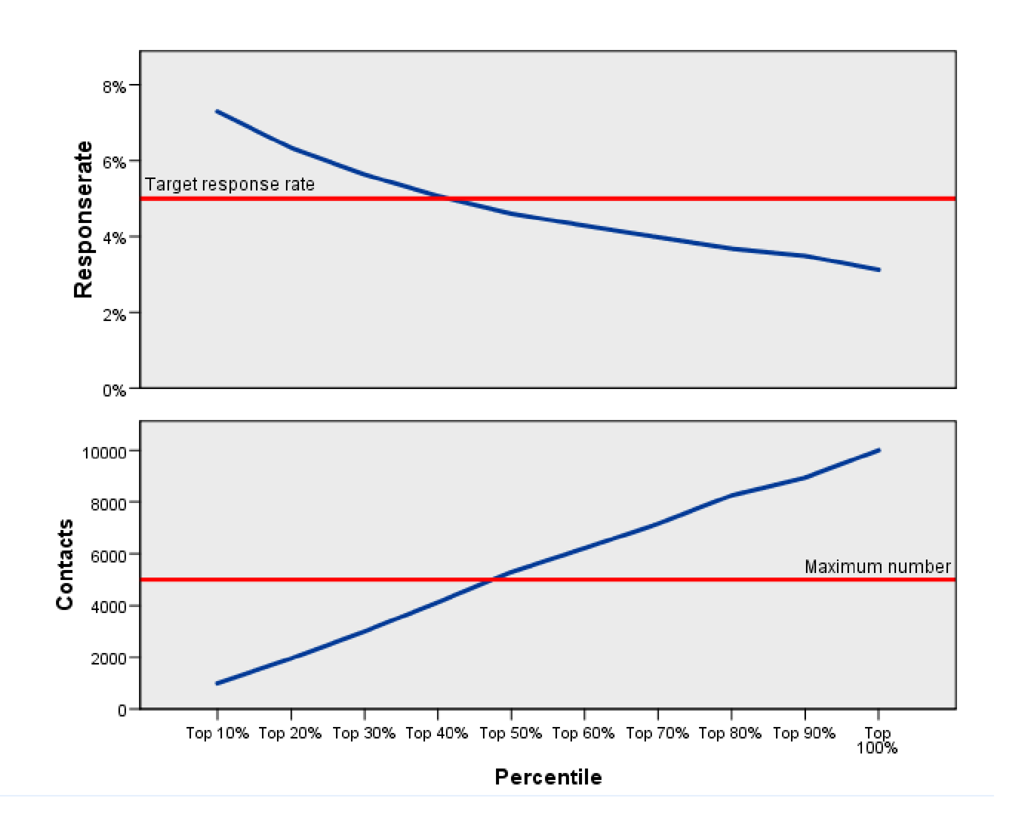
	PostalCode	ResponseRate	Responses	Contacts	Index	Rank	var
1	932	10.00%	4.00	40	3.60	1	
2	098	8.82%	6.00	68	5.47	1	
3	740	7.76%	9.00	116	8.30	1	
4	100	7.69%	7.00	91	6.46	1	
5	110	7.69%	5.00	65	4.62	1	
6	954	7.55%	4.00	53	3.70	1	
7	108	7.32%	6.00	82	5.56	1	
8	107	7.04%	5.00	71	4.65	1	
9	090	6.90%	4.00	58	3.72	1	
10	966	6.90%	4.00	58	3.72	1	
11	760	6.72%	8.00	119	7.46	1	
12	113	6.25%	5.00	80	4.69	1	
13	927	6.00%	3.00	50	2.82	1	
14	969	6.00%	3.00	50	2.82	1	

Abbildung 5-2
Tabelle und Diagramm mit Zusammenfassung

Responserate

Percentile	Responserate	Contacts	Cumulative Response Rate	Total Contacts
Top 10%	7.3	1001	7.3	1001
Top 20%	5.3	956	6.3	1957
Top 30%	4.3	1042	5.6	2999
Top 40%	3.5	1127	5.1	4126
Top 50%	3.0	1173	4.6	5299
Top 60%	2.4	914	4.3	6213
Top 70%	2.0	948	4.0	7161
Top 80%	1.7	1095	3.7	8256
Top 90%	1.2	680	3.5	8936
Top 100%	.0	1064	3.1	10000

Green Red Caption



Das neue Datenblatt enthält folgende Felder:

- **Postleitzahl.** Falls die Postleitzahl-Gruppen auf nur einem Teil des Gesamtwerts basieren, ist dies der Wert dieses Teils der Postleitzahl. Das Kopfzeilenlabel für diese Spalte in der Excel-Datei ist der Name des Postleitzahlfelds im ursprünglichen Daten-Set.
- **Responserate.** Der Prozentsatz der positiven Antworten in jeder Postleitzahl-Gruppe.
- **Antworten.** Der Anzahl der positiven Antworten in jeder Postleitzahl-Gruppe.

- **Kontakte.** Die Gesamtanzahl von Kontakten in jedem Postleitzahlbereich, die einen nicht fehlenden Wert für das Responsefeld enthalten.
- **Index.** Die "gewichtete" Antwort auf Basis der Formel *N x P x (1-P)*, wobei *N* die Anzahl von Kontakten und *P* die als Anteil ausgedrückte Responserate ist.
- **Rang.** Dezil-Rang (oberste 10 %, oberste 20 % usw.) der kumulativen Postleitzahl-Responseraten in absteigender Reihenfolge.

Erläuterung der Daten für Responseraten nach Postleitzahlen

Responsefeld. Das Responsefeld kann ein String-Feld oder ein numerisches Feld sein. Falls dieses Feld einen Wert enthält, der die Anzahl von Käufen oder ihren Geldwert anzeigt, müssen Sie ein neues Feld erstellen, in dem ein einzelner Wert sämtliche positiven Antworten repräsentiert. Für weitere Informationen siehe Erstellen eines kategorialen Responsefelds auf S. 33.

Wert für positive Antworten. Der Wert für positive Antworten bezeichnet diejenigen Kunden, die positiv reagiert haben (zum Beispiel, indem sie einen Kauf getätigt haben). Es wird davon ausgegangen, dass alle anderen nicht fehlenden Responsewerte eine negative Antwort anzeigen. Falls es definierte Wertelabels für das Responsefeld gibt, werden diese Labels in der Dropdown-Liste angezeigt.

Postleitzahlfeld. Das Postleitzahlfeld kann ein String-Feld oder ein numerisches Feld sein.

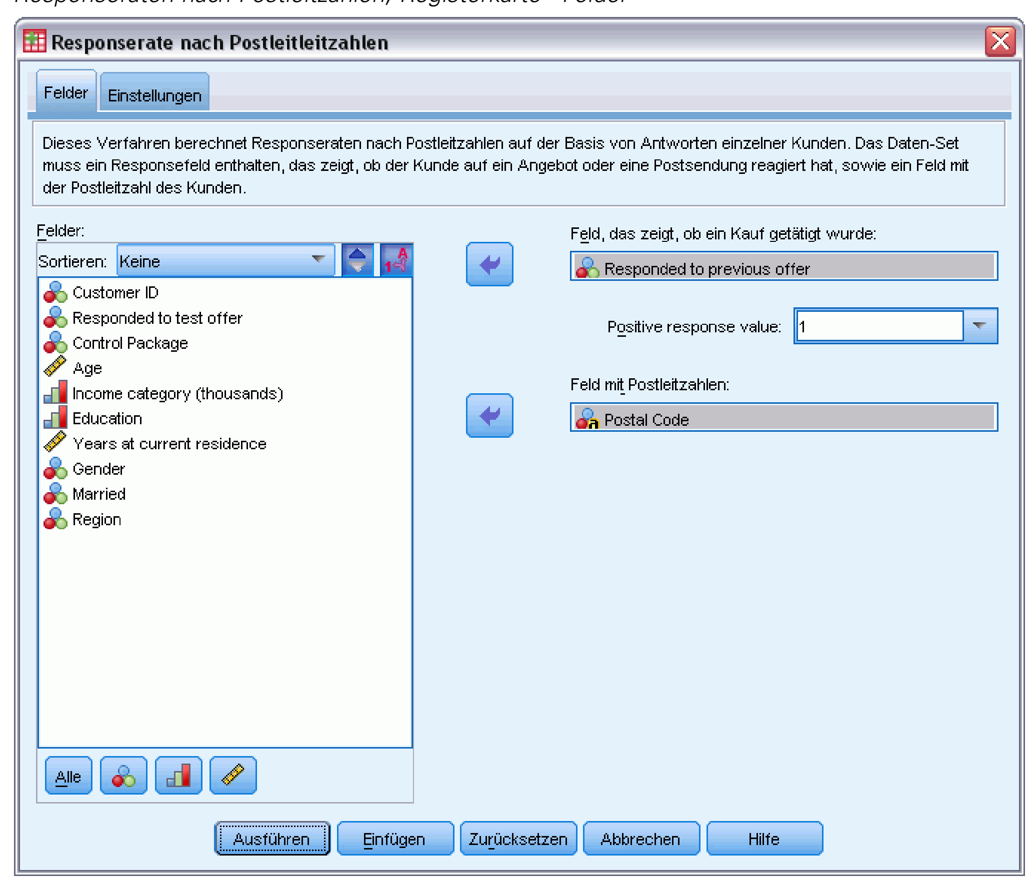
So erhalten Sie Responseraten nach Postleitzahlen

Wählen Sie die folgenden Befehle aus den Menüs aus:

Option "Direct Marketing" (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Postleitzahlbereiche mit den meisten Antworten identifizieren.

Abbildung 5-3
Responseraten nach Postleitzahlen, Registerkarte "Felder"



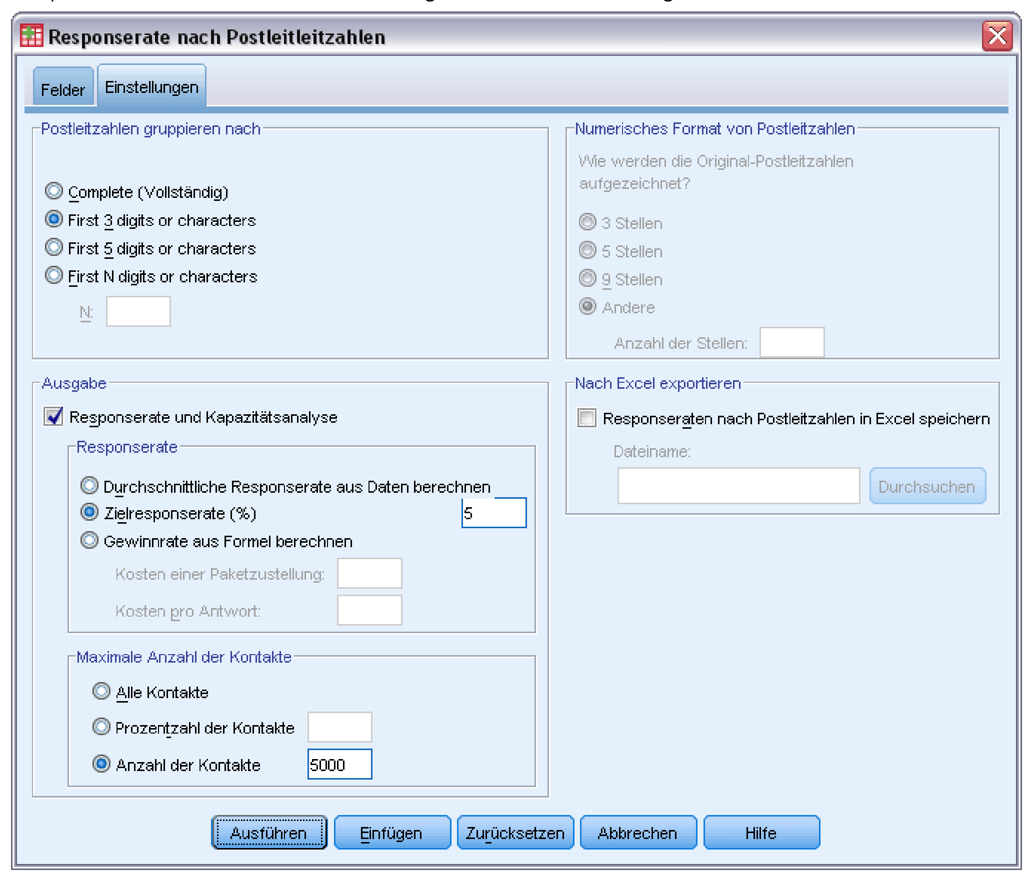
- ▶ Wählen Sie das Feld aus, das anzeigt, welche Kontakte auf das Angebot reagiert haben.
- ▶ Geben Sie den Wert an, der eine positive Antwort anzeigt. Für Werte mit definierten Wertelabels können Sie das Wertelabel aus der Dropdown-Liste auswählen, woraufhin der entsprechende Wert angezeigt wird.
- ▶ Wählen Sie das Feld, das die Postleitzahl enthält.
- ▶ Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Außerdem sind die folgenden Optionen verfügbar:

- Anstelle des vollständigen Werts können Sie Responseraten auf Basis der ersten *n* Zeichen oder Stellen der Postleitzahl generieren.
- Sie können die Ergebnisse automatisch als Excel-Datei speichern.
- Anzeigeoptionen der Ausgabe anpassen

Einstellungen

Abbildung 5-4
Responseraten nach Postleitzahlen, Registerkarte "Einstellungen"



Postleitzahlen gruppieren nach

Dadurch wird festgelegt, wie Datensätze gruppiert werden, um Responseraten zu berechnen. Standardmäßig wird dazu die gesamte Postleitzahl verwendet und alle Datensätze mit derselben Postleitzahl werden zur Berechnung der Gruppen-Responserate gruppiert. Alternativ können Sie Datensätze auch anhand eines Teils der vollständigen Postleitzahl gruppieren, welcher aus den ersten *n* Stellen oder Zeichen besteht. Dies ist nützlich, wenn Sie beispielsweise nur die ersten fünf Zeichen einer Postleitzahl aus zehn Zeichen oder die ersten drei Stellen einer fünfstelligen Postleitzahl für die Gruppierung verwenden möchten. Das Ausgabe-Daten-Set wird einen Datensatz für jede Postleitzahl-Gruppe enthalten. Falls Sie einen Wert eingeben, muss es sich dabei um eine positive ganze Zahl handeln.

Numerisches Format von Postleitzahlen

Wenn das Postleitzahlfeld numerisch ist und Sie die Postleitzahlen auf Basis der ersten *n* Stellen anstatt des Gesamtwerts gruppieren möchten, müssen Sie die Anzahl von Stellen des ursprünglichen Werts angeben. Die Anzahl von Stellen ist die *maximal* mögliche Anzahl von

Stellen der Postleitzahl. Falls das Postleitzahlenfeld beispielsweise sowohl fünfstellige als auch neunstellige Postleitzahlen enthält, sollten Sie als Anzahl von Stellen 9 eingeben.

Anmerkung: Abhängig vom Anzeigeformat werden manche fünfstelligen Postleitzahlen unter Umständen mit nur vier Stellen angezeigt, wobei aber eine führende Null impliziert ist.

Ausgabe

Neben dem neuen Daten-Set, das die Responseraten nach Postleitzahl enthält, können Sie auch eine Tabelle und ein Diagramm anzeigen, die die Ergebnisse nach Dezil-Rang zusammenfassen (oberste 10 %, oberste 20 % usw.). In der Tabelle werden Responseraten, kumulative Responseraten, die Anzahl von Datensätzen sowie die kumulative Anzahl von Datensätzen in jedem Dezil angezeigt. Im Diagramm werden kumulative Responseraten sowie die kumulative Anzahl von Datensätzen in jedem Dezil angezeigt.

Akzeptable Mindest-Responserate. Wenn Sie eine akzeptable Mindest-Responserate oder eine Break-Even-Formel eingeben, wird die Tabelle farbkodiert, so dass erkennbar ist, welche Dezile der Mindestanforderung an die kumulative Responserate entsprechen. Das Diagramm enthält eine Bezugslinie, die den Wert der Mindest-Responserate kenntlich macht.

- **Zielresponserate.** In Prozent ausgedrückte Responserate (Prozentsatz der positiven Antworten in jeder Postleitzahl-Gruppe). Der Wert muss größer als 0 und kleiner als 100 sein.
- **Gewinnrate aus Formel berechnen.** Berechnen Sie die minimale kumulative Responserate anhand dieser Formel: *(Kosten der Postsendung/Nettoertrag pro Antwort) x 100*. Beide Werte müssen positive Zahlen sein. Das Ergebnis sollte ein Wert größer 0 und kleiner als 100 sein. Falls die Kosten einer Postsendung beispielsweise 0,75 Euro und der Nettoertrag pro Antwort 56 Euro betragen, beträgt die Mindest-Responserate: (0,75/56) x 100 = 1,34%.

Maximale Anzahl an Kontakten. Wenn Sie eine maximale Anzahl von Kontakten angeben, wird die Tabelle farbkodiert, so dass erkennbar ist, welche Dezile die kumulative maximale Anzahl von Kontakten (Datensätzen) nicht übersteigen. Das Diagramm enthält eine Bezugslinie, die diesen Wert kenntlich macht.

- **Prozentzahl der Kontakte.** Das in Prozent ausgedrückte Maximum. Dies ist nützlich, wenn Sie beispielsweise die Dezile mit den höchsten Responseraten ermitteln möchten, die nicht mehr als 50 % aller Kontakte enthalten. Der Wert muss größer als 0 und kleiner als 100 sein.
- **Anzahl der Kontakte.** Das als Anzahl der Kontakte angegebene Maximum. Dies ist nützlich, wenn Sie beispielsweise nicht mehr als 10.000 Sendungen verschicken möchten; in diesem Fall würden Sie den Wert auf 10.000 festlegen. Der Wert muss eine positive ganze Zahl sein (ohne Gruppierungssymbole).

Wenn Sie sowohl eine minimale akzeptable Responserate als auch eine maximale Anzahl von Kontakten angeben, erfolgt die Farbkodierung der Tabelle abhängig davon, welche Bedingung als erste erfüllt wird.

Nach Excel exportieren

Bei dieser Prozedur wird automatisch ein neues Daten-Set erstellt, das Responseraten nach Postleitzahlen enthält. Jeder Datensatz (Zeile) im Daten-Set steht dabei für eine Postleitzahl. Sie können dieselben Informationen automatisch als Excel-Datei speichern. Sie wird im Format "Excel 97-2003" gespeichert.

Erstellen eines kategorialen Responsefelds

Das Responsefeld sollte kategorial sein, wobei ein Wert alle positiven Reaktionen darstellen sollte. Es wird davon ausgegangen, dass alle anderen nicht fehlenden Werte eine negative Antwort anzeigen. Falls das Responsefeld einen stetigen (metrischen) Wert enthält, beispielsweise die Anzahl oder den Geldwert der Käufe, müssen Sie ein neues Feld erstellen, das allen von Null abweichenden Responsewerten eine einzelne positive Antwort zuweist.

- Falls negative Antworten als "0" (nicht leer, was als fehlender Wert aufgefasst wird) aufgezeichnet werden, kann dies anhand der folgenden Formel berechnet werden:

  NewName=OldName>0,

  wobei *NewName* der Name des neuen Felds und *OldName* der Name des neuen Felds ist. Dies ist ein logischer Ausdruck, der allen nicht fehlenden Werten größer 0 einen Wert von 1 und allen nicht fehlenden Werten kleiner oder gleich 0 den Wert 0 zuweist.

- Falls für negative Antworten kein Wert aufgezeichnet wird, werden diese Werte als fehlend behandelt und die Formel ist etwas komplizierter:

  NewName=NOT(MISSING(OldName))

  Bei diesem logischen Ausdruck wird allen nicht fehlenden Responsewerten ein Wert von 1 und allen fehlenden Responsewerten ein Wert von 0 zugewiesen.

- Falls Sie zwischen negativen (0) Responsewerten und fehlenden Werten nicht unterscheiden können, kann kein korrekter Responsewert berechnet werden. Falls es nur relativ wenig tatsächlich fehlende Werte gibt, muss dies jedoch keine großen Auswirkungen auf die berechneten Responseraten haben. Falls es jedoch viele fehlende Werte gibt – z. B. wenn die Responseinformationen nur für eine kleine Teststichprobe des gesamten Daten-Sets berechnet werden –, wird dies dazu führen, dass die berechneten Responseraten bedeutungslos sind, da sie deutlich niedriger sein werden als die tatsächlichen Responseraten.

So erstellen Sie ein kategoriales Responsefeld

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Transformieren
  Variable berechnen

▶ Geben Sie für "Zielvariable" einen neuen Feld-(Variablen-)Namen ein.

▶ Falls negative Reaktionen als 0 aufgezeichnet werden, geben Sie als numerischen Ausdruck OldName>0 ein, wobei *OldName* der ursprüngliche Feldname ist.

▶ Falls negative Reaktionen als fehlend (leer) aufgezeichnet werden, geben Sie als numerischen Ausdruck NOT(MISSING(OldName)) ein, wobei *OldName* der ursprüngliche Feldname ist.

Kapitel
6
Kaufneigung

Für die Kaufneigung werden Ergebnisse einer Testsendung oder einer früheren Kampagne verwendet, um Bewertungen zu erstellen. Die Bewertungen zeigen an, bei welchen Kontakten die Wahrscheinlichkeit einer Antwort am höchsten ist. Das Responsefeld zeigt, wer auf die Testsendung oder die frühere Kampagne reagiert hat. Die Neigungsfelder sind die Merkmale, die Sie verwenden, um die Wahrscheinlichkeit einer Antwort seitens Kontakten mit ähnlichen Eigenschaften vorherzusagen.

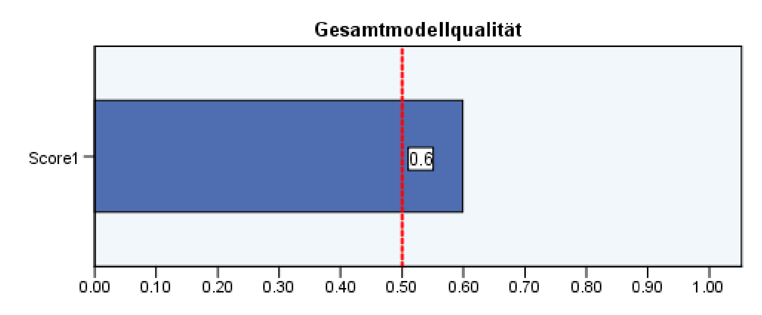
Bei dieser Technik wird die binäre logistische Regression verwendet, um die Neigungsbewertungen zu erstellen.

Beispiel. Die Direktmarketing-Abteilung eines Unternehmens verwendet die Ergebnisse einer Testsendung, um den übrigen Kontakten in ihrer Datenbank Neigungsbewertungen zuzuweisen, wobei verschiedene demographische Merkmale eingesetzt werden, um Kontakte zu ermitteln, bei denen die Wahrscheinlichkeit einer Antwort und eines Kaufs am größten ist.

Ausgabe

Bei dieser Prozedur wird automatisch ein neues Feld im Daten-Set erstellt, das Neigungsbewertungen enthält. In der optionalen Diagnosenausgabe sind ein Diagramm zur Gesamtmodellqualität sowie eine Klassifikationsmatrix enthalten, die vorhergesagte Antworten mit tatsächlichen Antworten vergleicht. Sie können außerdem ein neues Daten-Set oder eine Excel-Datei erstellen, die ausschließlich Kontakte enthalten, welche bestimmten Kriterien entsprechen (z.B. minimale Neigungsbewertung).

Abbildung 6-1
Diagramm zur Gesamtmodellqualität



Erläuterung der Daten zur Kaufneigung

Responsefeld. Das Responsefeld kann ein String-Feld oder ein numerisches Feld sein. Falls dieses Feld einen Wert enthält, der die Anzahl von Käufen oder ihren Geldwert anzeigt, müssen Sie ein neues Feld erstellen, in dem ein einzelner Wert sämtliche positiven Antworten repräsentiert. Für weitere Informationen siehe Erstellen eines kategorialen Responsefelds auf S. 42.

Wert für positive Antworten. Der Wert für positive Antworten bezeichnet diejenigen Kunden, die positiv reagiert haben (zum Beispiel, indem sie einen Kauf getätigt haben). Es wird davon ausgegangen, dass alle anderen nicht fehlenden Responsewerte eine negative Antwort anzeigen. Falls es definierte Wertelabels für das Responsefeld gibt, werden diese Labels in der Dropdown-Liste angezeigt.

Anmerkung: Bei der Prozedur werden allen Datensätzen innerhalb der Datendatei Bewertungen zugewiesen, unabhängig davon, ob sie in der Postsendungs-Kampagne enthalten waren oder nicht. Eine “negative Antwort” bedeutet, dass der Kontakt in der Postsendungs-Kampagne enthalten war, aber nicht positiv reagiert hat (zum Beispiel indem er keinen Kauf getätigt hat). Die Daten sollten zwischen Kontakten, die in der Postsendungs-Kampagne enthalten waren und nicht reagiert haben, und Kontakten, die nicht reagiert haben, weil sie nicht in der Kampagne enthalten waren, unterscheiden. Für weitere Informationen siehe Erstellen eines kategorialen Responsefelds auf S. 42.

Neigung vorhersagen durch. Die Felder, die verwendet werden, um die Neigung vorherzusagen, können String-Felder oder numerische Felder und außerdem nominal, ordinal oder stetig (metrisch) sein – es ist jedoch wichtig, allen Feldern für Einflussgrößen das geeignete Messniveau zuzuweisen.

Messniveau. Es ist wichtig, das korrekte Messniveau zuzuweisen, da sich dieses auf die Berechnung der Ergebnisse auswirkt.

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Kontinuierlich.** Eine Variable kann als metrisch (stetig) behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

Ein Symbol neben jedem Feld zeigt das aktuelle Messniveau an.

Messniveau	**Datentyp**			
	Numerisch	Zeichenfolge	Datum	Zeit

Metrisch (stetig)		entfällt		
Ordinal				
Nominal				

Sie können das Messniveau in der Variablenansicht des Daten-Editors ändern oder das Dialogfeld "Variableneigenschaften definieren" verwenden, um ein geeignetes Messniveau für jedes Feld anzugeben .

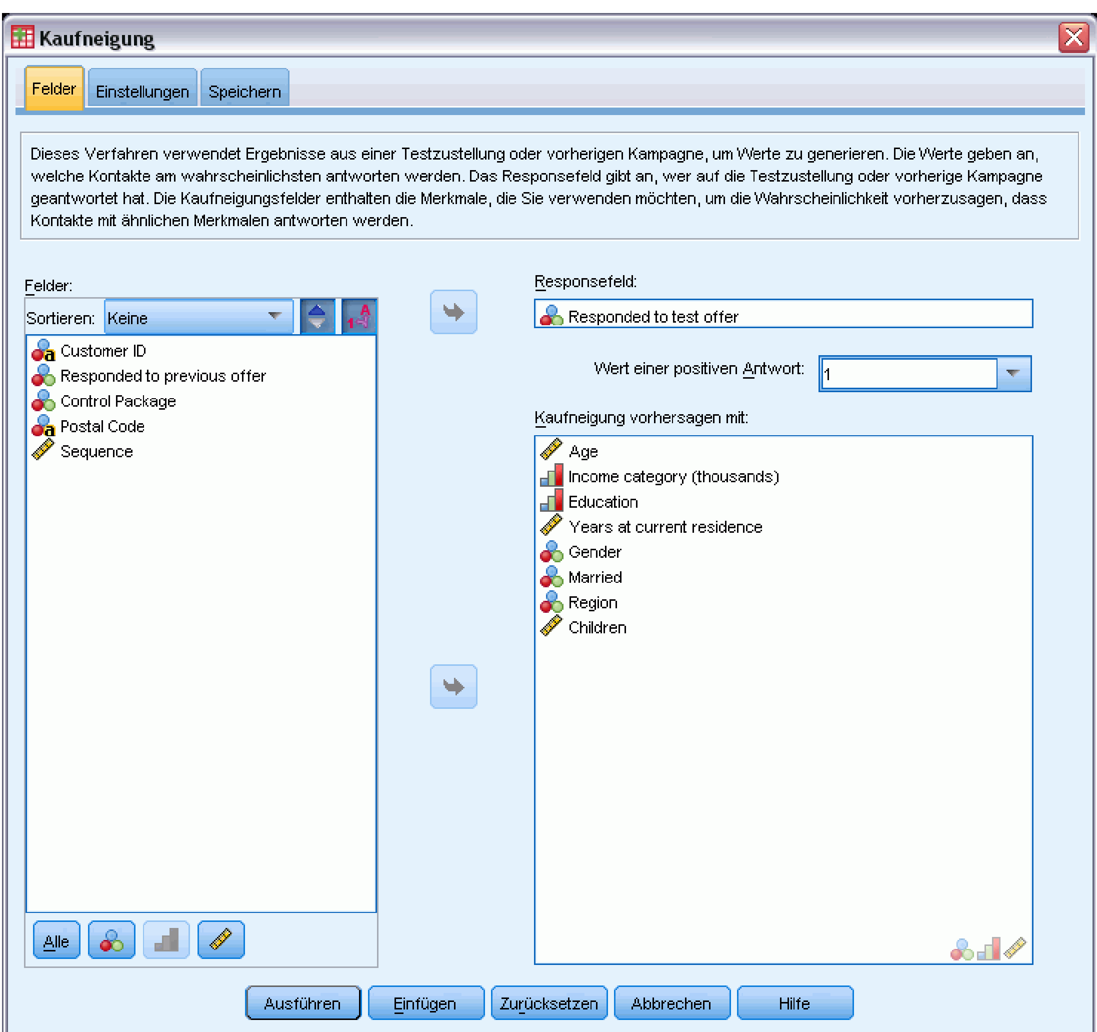
So erhalten Sie Kaufneigungsbewertungen

Wählen Sie die folgenden Befehle aus den Menüs aus:

Option "Direct Marketing" (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Kontakte mit höchster Kaufneigung auswählen.

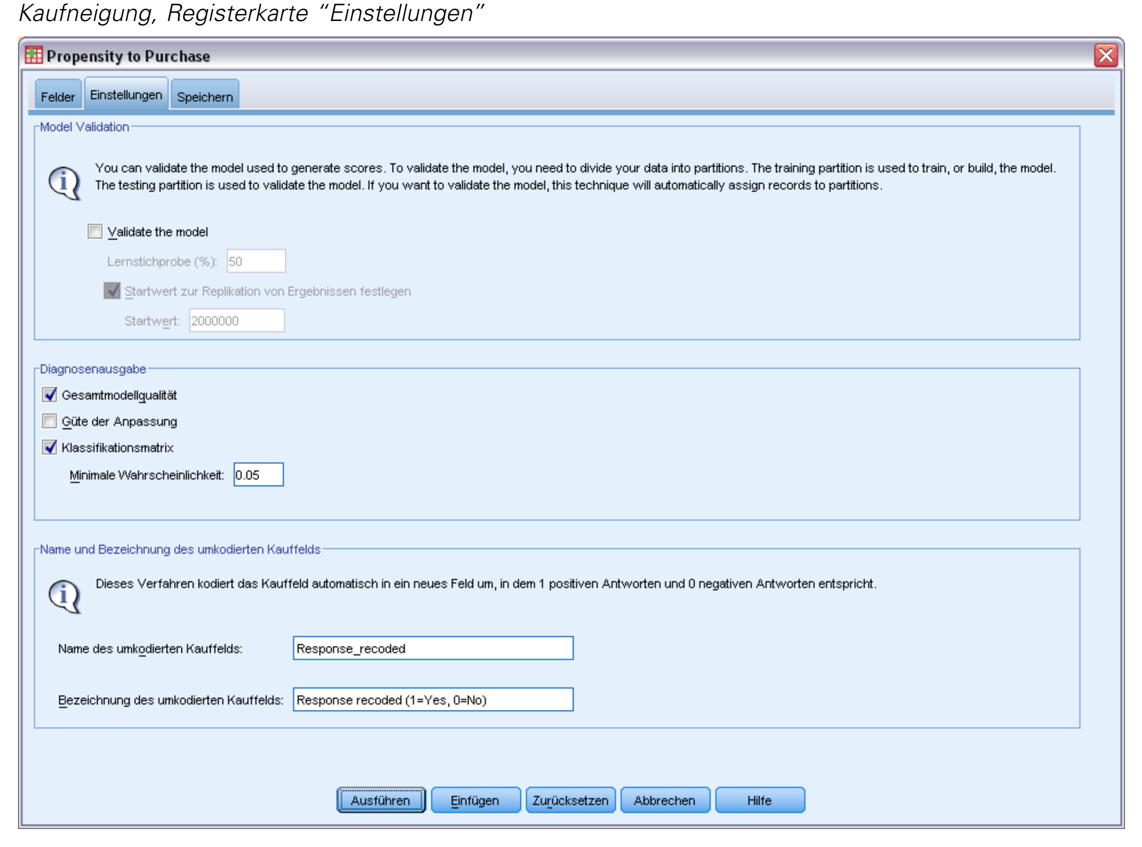
Abbildung 6-2
Kaufneigung, Registerkarte "Felder"



- Wählen Sie das Feld aus, das anzeigt, welche Kontakte auf das Angebot reagiert haben.
- Geben Sie den Wert an, der eine positive Antwort anzeigt. Für Werte mit definierten Wertelabels können Sie das Wertelabel aus der Dropdown-Liste auswählen, woraufhin der entsprechende Wert angezeigt wird.
- Wählen Sie die Felder aus, die Sie verwenden möchten, um die Neigung vorherzusagen.
- Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Einstellungen

Abbildung 6-3
Kaufneigung, Registerkarte "Einstellungen"



Modellvalidierung

Bei der Modellvalidierung werden zu Diagnosezwecken Trainings- und Testgruppen erstellt. Falls Sie die Klassifikationsmatrix im Abschnitt "Diagnoseausgabe" auswählen, wird die Tabelle zu Vergleichszwecken in (ausgewählte) Trainings- und (nicht ausgewählte) Testabschnitte unterteilt. Wählen Sie die Modellvalidierung nur aus, wenn Sie auch die Klassifikationsmatrix auswählen. Die Bewertungen erfolgen auf Basis des Modells, das aus der Trainings-Stichprobe erstellt wurde, deren Anzahl enthaltener Datensätze stets niedriger als die Gesamtanzahl verfügbarer Datensätze ist. Ein Beispiel: Die Standardgröße für Trainings-Stichproben ist 50 %, und ein Modell, das auf Basis der Hälfte aller verfügbaren Datensätze erstellt wird, kann nicht so zuverlässig sein, wie ein Modell auf Basis aller verfügbaren Datensätze.

- **Partitionsgröße der Lernstichprobe (%).** Legen Sie den Prozentsatz der Datensätze fest, die der Trainingsstichprobe zugewiesen werden sollen. Die übrigen Datensätze mit nicht fehlenden Werten für das Responsefeld werden der Teststichprobe zugewiesen. Der Wert muss größer als 0 und kleiner als 100 sein.
- **Startwert zur Replikation von Ergebnissen festlegen.** Da die Zuweisung von Datensätzen zu den Trainings- und Teststichproben auf Zufallsbasis geschieht, erhalten Sie unter Umständen bei jeder Durchführung der Prozedur unterschiedliche Ergebnisse, es sei denn, Sie geben jedes Mal denselben Startwert für Zufallszahlen an.

Diagnosenausgabe

Gesamtmodellqualität. Zeigt ein Balkendiagramm der Gesamtmodellqualität an, die als ein Wert zwischen 0 und 1 ausgedrückt wird. Ein gutes Modell sollte einen Wert größer 0,5 aufweisen.

Güte der Anpassung. Zeigt eine Tabelle an, die Informationen darüber bietet, wie gut das Modell an die Daten angepasst ist. Für eine gute Anpassung sollte der Signifikanzwert größer 0,05 sein.

Klassifikationsmatrix. Zeit eine Matrix an, die die vorhergesagten positiven und negativen Antworten mit den tatsächlichen positiven und negativen Antworten vergleicht. Die Gesamtgenauigkeitsrate kann Aufschluss darüber geben, die gut das Modell funktioniert, aber möglicherweise interessieren Sie sich mehr für den Prozentsatz korrekt vorhergesagter positiver Antworten.

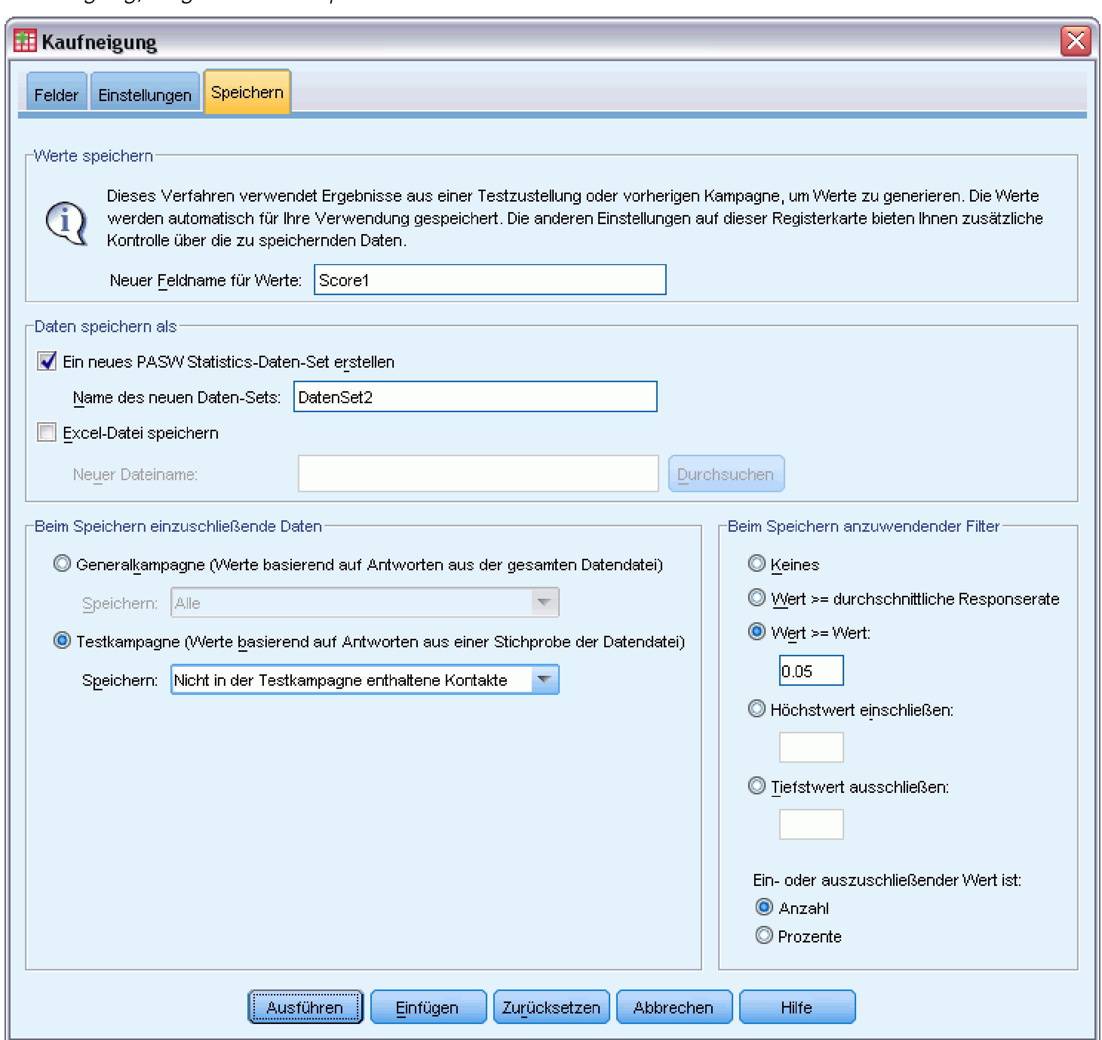
- **Minimale Wahrscheinlichkeit.** Weist der Kategorie für vorhergesagte positive Antworten in der Klassifikationsmatrix Datensätze mit einem Bewertungswert zu, der höher als der angegebene Wert ist. Die Bewertungen, die durch die Prozedur erstellt werden, stehen für die Wahrscheinlichkeit, dass der Kontakt positiv reagieren wird (zum Beispiel indem er einen Kauf tätigt). Allgemein sollten Sie einen Wert angeben, der in der Nähe Ihrer minimalen, als Anteil ausgedrückten Zielresponserate liegt. Falls Sie zum Beispiel an einer Responserate von mindestens 5 % interessiert sind, geben Sie 0,05 an. Der Wert muss größer als 0 und kleiner als 1 sein.

Name und Beschriftung des umkodierten Responsefelds

Dieses Verfahren kodiert das Responsefeld automatisch in ein neues Feld um, in dem “1” positiven Antworten und “0” negativen Antworten entspricht. Die Analyse wird für das umkodierte Feld durchgeführt. Sie können den Standardnamen und die Standardbeschriftung durch eigene Angaben ersetzen. Die Namen müssen den Benennungsregeln von PASW Statistics entsprechen.

Speichern (Kaufneigung)

Abbildung 6-4
Kaufneigung, Registerkarte "Speichern"



Werte speichern

Im ursprünglichen Daten-Set wird automatisch ein neues Feld mit Neigungsbewertungen gespeichert. Die Bewertungen stehen für die Wahrscheinlichkeit einer positiven Antwort, welche als Anteil ausgedrückt wird.

- Die Feldnamen müssen den Benennungsregeln von PASW Statistics entsprechen.
- Der Feldname kann kein Duplikat eines Feldnamens sein, der bereits im Daten-Set vorhanden ist. Falls Sie diese Prozedur also mehr als einmal mit demselben Daten-Set ausführen, müssen Sie jedes Mal einen anderen Namen angeben.

Daten speichern als

Neues PASW Statistics-Daten-Set erstellen. Erstellt ein neues Daten-Set in der aktuellen Sitzung. Dieses Daten-Set enthält alle Felder des ursprünglichen Daten-Sets, darunter auch das neue Neigungsbewertungsfeld. Dies ist vor allem nützlich, wenn Sie eine Untergruppe von Kontakten auf Basis der Einstellungen in den Optionen “Einzuschließende Daten” und/oder “Anzuwendender Filter” erstellen; anderenfalls ist es einfach eine exakte Kopie des ursprünglichen Daten-Sets. Wenn Sie Angaben für die Optionen “Einzuschließende Daten” und/oder “Anzuwendender Filter” machen, enthält das Daten-Set ausschließlich die Datensätze (Kontakte), die den angegebenen Kriterien entsprechen.

Als Excel-Datei speichern. Speichert alle Felder des ursprünglichen Daten-Sets, darunter auch das neue Neigungsbewertungsfeld, im Format “Excel 97-2003”. Wenn Sie Angaben für die Optionen “Einzuschließende Daten” und/oder “Anzuwendender Filter” machen, enthält die Excel-Datei ausschließlich die Datensätze (Kontakte), die den angegebenen Kriterien entsprechen.

Beim Speichern einzuschließende Daten

Pauschalkampagne. Falls die Bewertungen auf Antworten aus der gesamten Datendatei (z.B. einer früheren Kampagne) basieren, können Sie alle Kontakte, nur diejenigen, die positiv geantwortet haben (“Ja”), oder nur diejenigen, die negativ geantwortet haben (“Nein”), auswählen.

Testkampagne. Falls die Bewertungen auf einer Stichprobe aus der gesamten Datendatei (z.B. eine Testzustellung) basieren, können Sie die Kontakte, die nicht in der Testkampagne enthalten waren, nur diejenigen, die positiv geantwortet haben (“Ja”), oder nur diejenigen, die negativ geantwortet haben (“Nein”), auswählen.

Anmerkung: Bei der Prozedur werden allen Datensätzen innerhalb der Datendatei Bewertungen zugewiesen, unabhängig davon, ob sie in der Postsendungs-Kampagne enthalten waren oder nicht. Eine “negative Antwort” bedeutet, dass der Kontakt in der Postsendungs-Kampagne enthalten war, aber nicht positiv reagiert hat (zum Beispiel indem er keinen Kauf getätigt hat). Die Daten sollten zwischen Kontakten, die in der Postsendungs-Kampagne enthalten waren und nicht reagiert haben, und Kontakten, die nicht reagiert haben, weil sie nicht in der Kampagne enthalten waren, unterscheiden.

Beim Speichern anzuwendender Filter

Beim Filtern werden zusätzliche Auswahlkriterien angewandt, die über die Auswahlkriterien hinausgehen, die Sie möglicherweise in der Gruppe “Beim Speichern einzuschließende Daten” ausgewählt haben.

Keine. Es wird kein Auswahlfilter angewandt.

Bewertung >= durchschnittliche Responserate. Auf diese Weise können Sie nur Kontakte einschließen, deren Neigungsbewertung größer oder gleich der gesamten durchschnittlichen Responserate ist. Falls beispielsweise die Gesamtresponserate 1,5 % beträgt, werden nur Kontakte mit einer Neigungsbewertung von 0,015 oder höher eingeschlossen.

Bewertung >= Wert. Auf diese Weise können Sie nur Kontakte einschließen, deren Neigungsbewertung größer oder gleich dem angegebenen Wert ist. Die Bewertungen sind vorhergesagte Wahrscheinlichkeiten einer positiven Antwort, welche als Anteil ausgedrückt werden. Der Wert muss größer als 0 und kleiner als 1 sein.

Oberste einschließen. Auf diese Weise werden nur die obersten *n* oder die obersten *n%* der Kontakte mit den höchsten Neigungsbewertungen eingeschlossen. Falls Sie "Anzahl" auswählen, muss es sich beim Wert um eine positive ganze Zahl handeln. Falls Sie "Prozent" auswählen, muss der Wert größer als 0 und kleiner als 100 sein.

Unterste ausschließen. Auf diese Weise werden die untersten *n* oder die untersten *n%* der Kontakte mit den niedrigsten Neigungsbewertungen ausgeschlossen. Falls Sie "Anzahl" auswählen, muss es sich beim Wert um eine positive ganze Zahl handeln. Falls Sie "Prozent" auswählen, muss der Wert größer als 0 und kleiner als 100 sein.

Erstellen eines kategorialen Responsefelds

Das Responsefeld sollte kategorial sein, wobei ein Wert alle positiven Reaktionen darstellen sollte. Es wird davon ausgegangen, dass alle anderen nicht fehlenden Werte eine negative Antwort anzeigen. Falls das Responsefeld einen stetigen (metrischen) Wert enthält, beispielsweise die Anzahl oder den Geldwert der Käufe, müssen Sie ein neues Feld erstellen, das allen von Null abweichenden Responsewerten eine einzelne positive Antwort zuweist.

- Falls negative Antworten als "0" (nicht leer, was als fehlender Wert aufgefasst wird) aufgezeichnet werden, kann dies anhand der folgenden Formel berechnet werden:

  NewName=OldName>0,

  wobei *NewName* der Name des neuen Felds und *OldName* der Name des neuen Felds ist. Dies ist ein logischer Ausdruck, der allen nicht fehlenden Werten größer 0 einen Wert von 1 und allen nicht fehlenden Werten kleiner oder gleich 0 den Wert 0 zuweist.

- Falls für negative Antworten kein Wert aufgezeichnet wird, werden diese Werte als fehlend behandelt und die Formel ist etwas komplizierter:

  NewName=NOT(MISSING(OldName))

  Bei diesem logischen Ausdruck wird allen nicht fehlenden Responsewerten ein Wert von 1 und allen fehlenden Responsewerten ein Wert von 0 zugewiesen.

- Falls Sie zwischen negativen (0) Responsewerten und fehlenden Werten nicht unterscheiden können, kann kein korrekter Responsewert berechnet werden. Falls es nur relativ wenig tatsächlich fehlende Werte gibt, muss dies jedoch keine großen Auswirkungen auf die berechneten Responseraten haben. Falls es jedoch viele fehlende Werte gibt – z. B. wenn die Responseinformationen nur für eine kleine Teststichprobe des gesamten Daten-Sets berechnet werden –, wird dies dazu führen, dass die berechneten Responseraten bedeutungslos sind, da sie deutlich niedriger sein werden als die tatsächlichen Responseraten.

So erstellen Sie ein kategoriales Responsefeld

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Transformieren
  Variable berechnen

- Geben Sie für “Zielvariable” einen neuen Feld-(Variablen-)Namen ein.
- Falls negative Reaktionen als 0 aufgezeichnet werden, geben Sie als numerischen Ausdruck OldName>0 ein, wobei *OldName* der ursprüngliche Feldname ist.
- Falls negative Reaktionen als fehlend (leer) aufgezeichnet werden, geben Sie als numerischen Ausdruck NOT(MISSING(OldName)) ein, wobei *OldName* der ursprüngliche Feldname ist.

Kapitel

7

Kontrollpakettest

Dieses Verfahren vergleicht Marketingkampagnen, um herauszufinden, ob es hinsichtlich der Wirksamkeit signifikante Unterschiede zwischen verschiedenen Paketen oder Angeboten gibt. Die Kampagnenwirksamkeit wird anhand von Antworten gemessen. Das Kampagnenfeld identifiziert unterschiedliche Kampagnen, zum Beispiel Angebot A und Angebot B. Das Responsefeld zeigt an, wenn ein Kontakt auf die Kampagne geantwortet hat. Wählen Sie “Kaufbetrag” aus, wenn die Antwort als Kaufbetrag aufgezeichnet wird, zum Beispiel “99.99”. Wählen Sie “Antwort” aus, wenn die Antwort nur angibt, ob der Kontakt positiv reagiert hat oder nicht, zum Beispiel “Ja” oder “Nein”.

Beispiel. Die Direktmarketing-Abteilung eines Unternehmens möchte herausfinden, ob eine neue Verpackungsgestaltung mehr positive Antworten erzeugt als die bestehende Verpackung. Daher verschicken sie Testsendungen, um zu ermitteln, ob die neue Verpackung eine deutlich höhere positive Responserate erzeugt. Die Testsendung besteht aus einer Kontrollgruppe, die die aktuelle Verpackung erhält, und einer Testgruppe, an die die neue Verpackungsgestaltung geschickt wird. Die Ergebnisse der zwei Gruppen werden dann miteinander verglichen, um zu sehen, ob ein deutlicher Unterschied besteht.

Ausgabe

Die Ausgabe enthält eine Tabelle, in der Häufigkeiten und Prozentwerte von positiven und negativen Antworten für jede anhand des Kampagnenfelds definierte Gruppe sowie eine Tabelle, in der festgehalten wird, welche Gruppen stark voneinander abweichen.

Abbildung 7-1
Ausgabe des Kontrollpakettests

		Control Package			
		Control		Test	
		Anzahl	Anzahl der Spalten (%)	Anzahl	Anzahl der Spalten (%)
Effectivenss(1=YES 0=NO)	0	875	96.2%	945	93.8%
	1	35	3.8%	62	6.2%

Vergleiche der Spaltenanteile[a]

		Control Package	
		Control	Test
		(A)	(B)
Effectivenss(1=YES 0=NO)	0	B	
	1		A

Die Ergebnisse beruhen auf zweiseitigen Tests mit einem Signifikanzniveau von 0.05. Für jedes signifikante Paar wird die Erläuterung der Kategorie mit den kleineren Spaltenanteilen unter der Kategorie mit den größeren Spaltenanteilen angezeigt.

a.Mit Hilfe der Bonferroni-Korrektur werden die Tests an alle paarweisen Vergleiche innerhalb einer Zeile der jeweils innersten Untertabelle angepaßt.

Erläuterungen und Annahmen der Daten des Kontrollpakettests

Kampagnenfeld. Das Kampagnenfeld sollte kategorial (nominal oder ordinal) sein.

Wirksamkeits-Responsefeld. Wenn Sie für das Wirksamkeitsfeld “Kaufbetrag” auswählen, muss das Feld numerisch sein und das Messniveau sollte stetig (metrisch) sein.

Falls Sie nicht zwischen negativen (für den Kaufbetrag ein Wert von 0) Responsewerten und fehlenden Werten unterscheiden können, kann keine korrekte Responserate berechnet werden. Falls es nur relativ wenig tatsächlich fehlende Werte gibt, muss dies jedoch keine großen Auswirkungen auf die berechneten Responseraten haben. Falls es jedoch viele fehlende Werte gibt – z. B. wenn die Responseinformationen nur für eine kleine Teststichprobe des gesamten Daten-Sets berechnet werden –, wird dies dazu führen, dass die berechneten Responseraten bedeutungslos sind, da sie deutlich niedriger sein werden als die tatsächlichen Responseraten.

Annahmen. Diese Prozedur geht davon aus, dass jeder Kampagnengruppe zufällig Kontakte zugewiesen wurden. Anders ausgedrückt besteht keine spezielle Gruppenzuweisung hinsichtlich Demografie, Kaufverlauf oder anderen Merkmalen und bei allen Kontakten ist die Wahrscheinlichkeit, einer beliebigen Gruppe zugewiesen zu werden, gleich hoch.

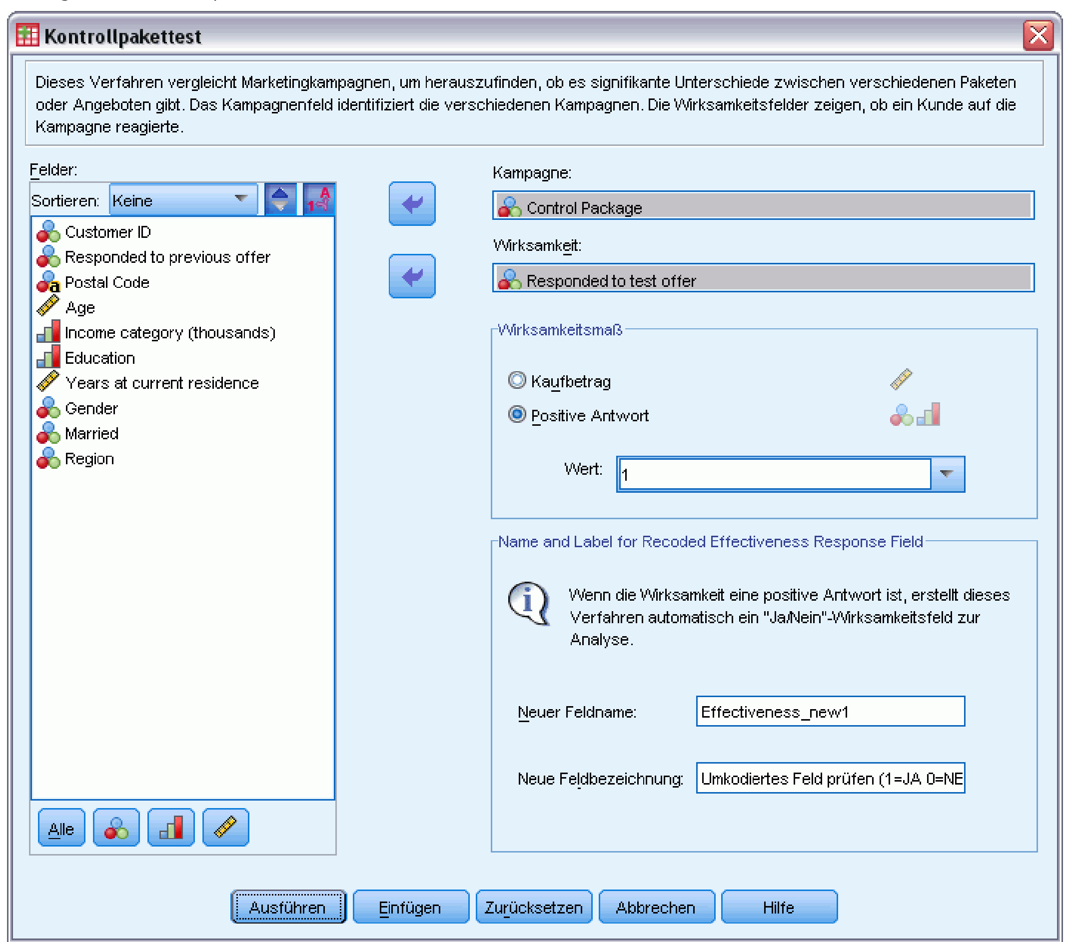
So führen Sie einen Kontrollpakettest durch

Wählen Sie die folgenden Befehle aus den Menüs aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

► Wählen Sie Wirksamkeit der Kampagnen vergleichen aus.

Abbildung 7-2
Dialogfeld "Kontrollpakettest"



- Wählen Sie das Feld aus, das anzeigt, zu welcher Kampagnengruppe jeder Kontakt gehört (zum Beispiel Angebot A, Angebot B etc.). Dieses Feld muss nominal oder ordinal sein.
- Wählen Sie das Feld aus, das die Responsewirksamkeit anzeigt.

Wenn das Responsefeld ein Kaufbetrag ist, muss das Feld numerisch sein.

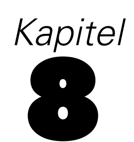
Wählen Sie Antwort aus, wenn das Responsefeld nur angibt, ob der Kontakt positiv reagiert hat oder nicht (zum Beispiel "Ja" oder "Nein"), und geben Sie den Wert ein, der eine positive Antwort darstellt. Für Werte mit definierten Wertelabels können Sie das Wertelabel aus der Dropdown-Liste auswählen, woraufhin der entsprechende Wert angezeigt wird.

Es wird automatisch ein neues Feld erstellt, in dem 1 positiven Antworten und 0 negativen Antworten entspricht; die Analyse wird in dem neuen Feld durchgeführt. Sie können den Standardnamen und die Standardbeschriftung durch eigene Angaben ersetzen. Die Namen müssen den Benennungsregeln von PASW Statistics entsprechen.

- Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Teil II: Beispiele

Kapitel



RFM-Analyse aus Transaktionsdaten

In einer Transaktionsdatei stellt jede Zeile eine eigene Transaktion anstelle eines eigenen Kunden dar. Es kann mehrere Transaktionszeilen für jeden Kunden geben. Dieses Beispiel verwendet die Datendatei *rfm_transactions.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

Transaktionsdaten

Das Daten-Set muss Variablen enthalten, die die folgenden Informationen enthalten:

- Eine Variable oder eine Kombination von Variablen, die jeden Fall (Kunden) identifizieren
- Eine Variable mit dem Datum jeder Transaktion
- Eine Variable mit dem Geldwert jeder Transaktion

Abbildung 8-1
RFM-Transaktionsdaten

ID	Date	Amount
1	08/04/2005	129
1	10/25/2004	50
1	07/24/2004	118
1	07/24/2004	136
1	09/04/2006	52
2	09/23/2005	183
2	11/05/2004	24
2	11/10/2005	66
2	12/03/2004	77
3	06/04/2005	102
3	05/15/2005	131

Durchführen der Analyse

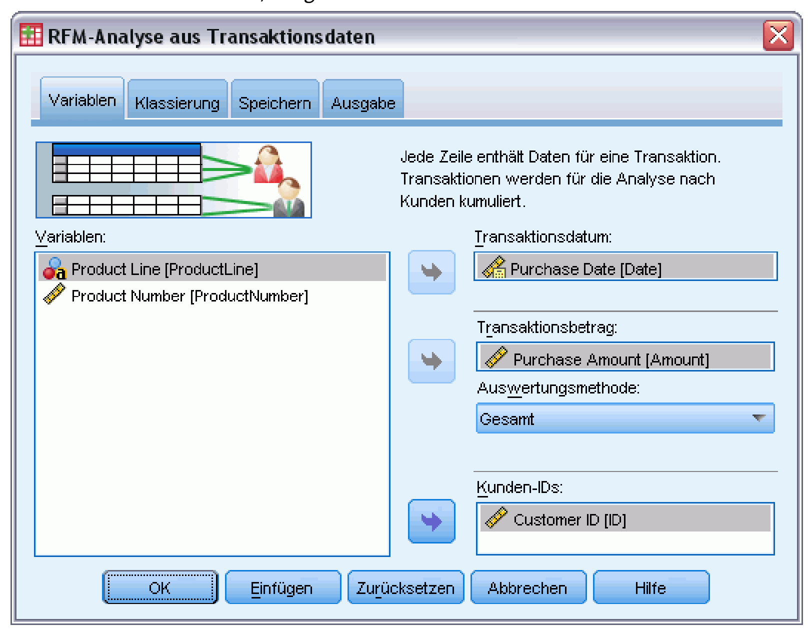
▶ Um Responseraten nach Postleitzahlen zu berechnen, wählen Sie in den Menüs folgende Optionen aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Hilfe bei der Erkennung meiner besten Kontakte (RFM-Analyse) und klicken Sie auf Weiter.

▶ Klicken Sie auf Transaktionsdaten und anschließend auf Weiter.

Abbildung 8-2
RFM aus Transaktionen, Registerkarte "Variablen"



▶ Klicken Sie auf Zurücksetzen, um frühere Einstellungen zu löschen.

▶ Wählen Sie für das Transaktionsdatum *Kaufdatum [Datum]*.

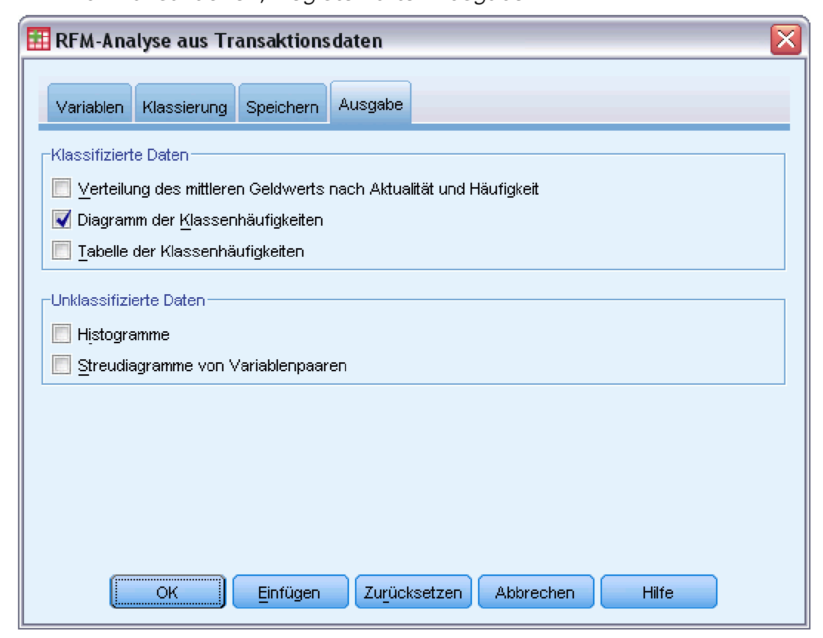
▶ Wählen Sie für den Transaktionsbetrag *Kaufbetrag [Betrag]*.

▶ Wählen Sie für die Auswertungsmethode Insgesamt.

▶ Wählen Sie für "Kunden-ID" *Kunden-ID [ID]*.

▶ Klicken Sie anschließend auf die Registerkarte Ausgabe.

Abbildung 8-3
RFM für Transaktionen, Registerkarte "Ausgabe"



▶ Wählen (markieren) Sie Diagramm der Klassenhäufigkeiten.

▶ Klicken Sie dann auf OK, um die Prozedur auszuführen.

Bewerten der Ergebnisse

Wenn Sie RFM-Scores aus Transaktionsdaten berechnen, wird ein neues Daten-Set erstellt, das die neuen RFM-Scores enthält.

Abbildung 8-4
RFM aus Daten-Set "Transaktionen"

ID	Aktuellstes_ Datum	Transaktion_ Häufigkeit	Betrag	Aktualitäts_ Score	Häufigkeits_ Score	Geldwert_ Score	RFM_Score
1	04-Sep-2006	5	485,00	4	3	4	434
2	10-Nov-2005	4	350,00	2	2	2	222
3	04-Jun-2005	2	233,00	1	2	4	124
4	18-Aug-2006	7	936,00	4	4	5	445
5	07-Jul-2006	3	359,00	4	1	5	415
6	16-Jul-2006	3	249,00	4	1	4	414
7	15-Feb-2006	7	1089,00	2	5	5	255

Standardmäßig enthält das Daten-Set die folgenden Informationen für jeden Kunden:

- Kunden-ID-Variable(n)
- Datum der letzten Transaktion
- Gesamtzahl der Transaktionen

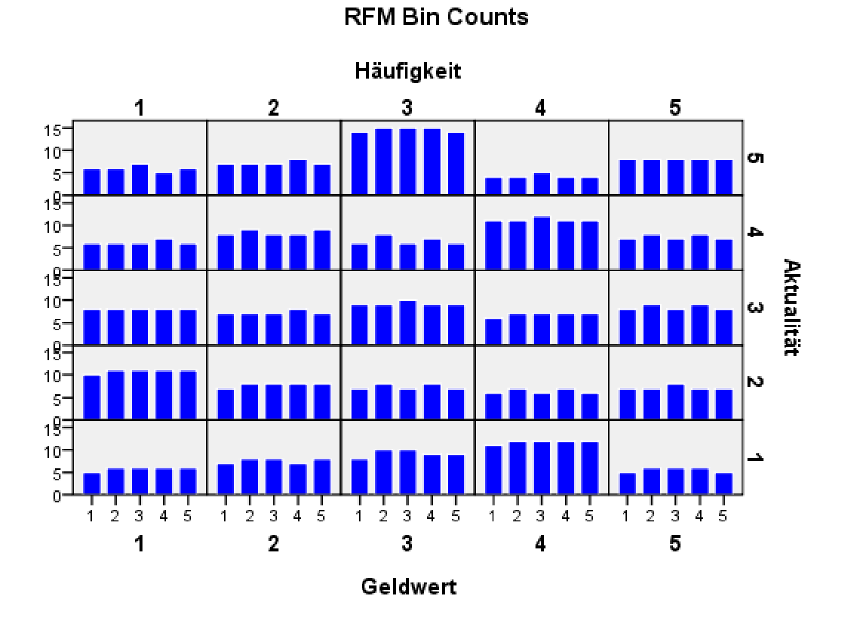
- Auswertung Transaktionsbetrag (Standard ist “Insgesamt”)
- Aktualität, Häufigkeit, Geldwert und kombinierte RFM-Scores

Das neue Daten-Set enthält nur eine Zeile (Datensatz) für jeden Kunden. Die Originaltransaktionsdaten wurden durch die Werte der Kunden-ID-Variablen aggregiert. Die ID-Variablen sind stets in dem neuen Daten-Set enthalten; anderenfalls hätten Sie keine Möglichkeit, die RFM-Scores den Kunden zuzuordnen.

Der kombinierte RFM-Score für jeden Kunden ist einfach die Konkatenation der drei einzelnen Scores berechnet als: (Aktualität x 100) + (Häufigkeit x 10) + Geldwert.

Das Diagramm der Klassenhäufigkeiten, das im Viewer-Fenster angezeigt wird, zeigt die Anzahl der Kunden in jeder RFM-Kategorie an.

Abbildung 8-5
Diagramm der Klassenhäufigkeiten



Die Standardmethode von fünf Score-Kategorien für jede der drei RFM-Komponenten führt zu 125 möglichen RFM-Score-Kategorien. Jeder Balken im Diagramm stellt die Anzahl der Kunden in jeder RFM-Kategorie dar.

Idealerweise wünschen Sie sich eine relativ gleichmäßige Verteilung der Kunden über alle RFM-Score-Kategorien. In der Realität tritt in der Regel eine gewisse Variation wie in diesem Beispiel auf. Wenn es viele leere Kategorien gibt, sollten Sie in Erwägung ziehen, die Klassifizierungsmethode zu ändern.

Es gibt eine Reihe von Strategien für den Umgang mit ungleichmäßigen Verteilungen von RFM-Scores wie:

- Verwendung verschachtelter anstelle von unabhängiger Klassifizierung
- Verringerung der Anzahl möglicher Score-Kategorien (Klassen)
- Wenn es eine große Anzahl an gebundenen Werten gibt, ordnen Sie Fälle mit den gleichen Scores zufällig unterschiedlichen Kategorien zu.

Für weitere Informationen siehe RFM-Klassifizierung in Kapitel 2 auf S. 6.

Kombinieren von Score-Daten mit Kundendaten

Nachdem Sie jetzt über ein Daten-Set verfügen, das RFM-Scores enthält, müssen Sie diese Scores den Kunden zuordnen. Sie könnten die Scores zurück in die Transaktionsdatendatei einfließen lassen, typischerweise wollen Sie die Score-Daten aber mit einer Datendatei kombinieren, die wie das RFM-Score-Daten-Set eine Zeile (Datensatz) für jeden Kunden – und auch Informationen wie Kundenname und Adresse – enthält.

Abbildung 8-6
RFM-Score-Daten-Set in der Variablenansicht

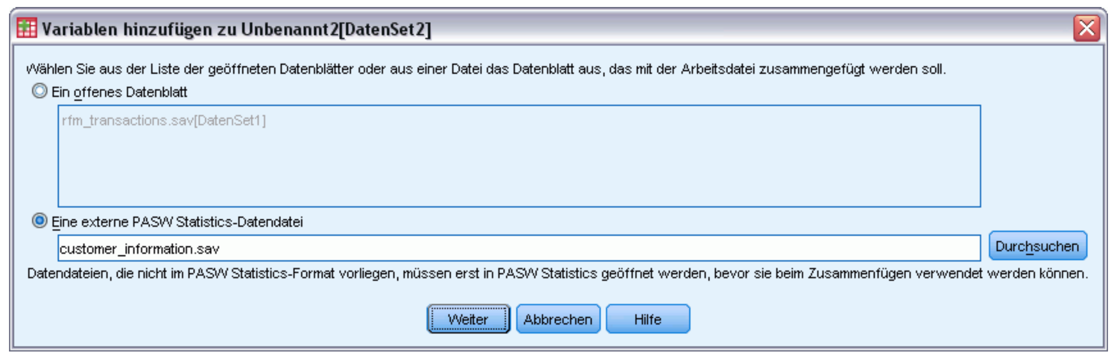
Name	Typ	Spaltenformat	Dezimalstellen	Variablenlabel	Wertelab
ID	Numerisch	5	0	Customer ID	Keine
Aktuellstes_Datum	Datum	10	0	Datum der letzt...	Keine
Transaktion_Häufigkeit	Numerisch	7	0	Anzahl an Tran...	Keine
Betrag	Numerisch	8	2	Betrag	Keine
Aktualitäts_Score	Numerisch	3	0	Aktualitäts-Score	Keine
Häufigkeits_Score	Numerisch	3	0	Häufigkeits-Score	Keine
Geldwert_Score	Numerisch	3	0	Geldwert-Score	Keine
RFM_Score	Numerisch	3	0	RFM-Score	Keine

- ▶ Machen Sie das Daten-Set, das die RFM-Scores enthält, zum aktiven Daten-Set. (Klicken Sie an eine beliebige Stelle im Fenster "Daten-Editor" eines Daten-Sets.)

- ▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
  Dateien zusammenfügen
    Variablen hinzufügen

Abbildung 8-7
Variablen hinzufügen, Dialogfeld "Dateien auswählen"

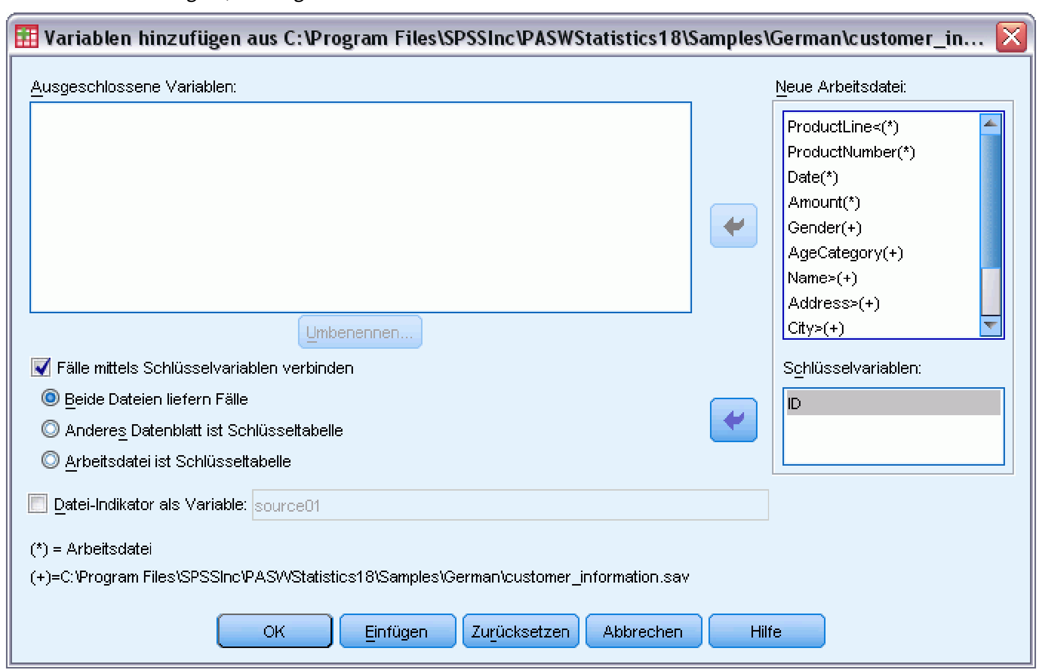


- ▶ Wählen Sie Externe Datendatei.

- ▶ Verwenden Sie die Schaltfläche Durchsuchen, um zum Ordner *Samples* zu wechseln, und wählen Sie *customer_information.sav* aus. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

▶ Klicken Sie dann auf Weiter.

Abbildung 8-8
Variablen hinzufügen, Dialogfeld "Variablen auswählen"



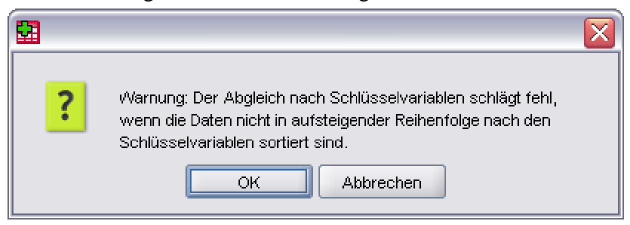
▶ Wählen (aktivieren) Sie Fälle mittels Schlüsselvariablen verbinden.

▶ Wählen Sie Beide Dateien liefern Fälle.

▶ Wählen Sie *ID* für die Liste "Schlüsselvariablen".

▶ Klicken Sie auf OK.

Abbildung 8-9
Warnmeldung "Variablen hinzufügen"



Achten Sie auf die Meldung, die Sie darauf hinweist, dass beide Dateien in aufsteigender Reihenfolge der Schlüsselvariablen sortiert sein müssen. In diesem Beispiel sind beide Dateien bereits in aufsteigender Reihenfolge der Schlüsselvariablen (die Kunden-ID-Variable, die bei der Berechnung der RFM-Scores ausgewählt wurde) sortiert. Wenn Sie RFM-Scores aus Transaktionsdaten berechnen, wird das neue Daten-Set automatisch in aufsteigender Reihenfolge der Kunden-ID-Variablen sortiert. Wenn Sie die Sortierfolge des Score-Daten-Sets ändern oder die Datendatei, mit der Sie das Score-Daten-Set zusammenfügen, nicht in dieser Reihenfolge sortiert

ist, müssen Sie zuerst beide Dateien in aufsteigender Reihenfolge der Kunden-ID-Variablen sortieren.

▶ Klicken Sie auf OK, um die beiden Daten-Sets zusammenzufügen.

Das Daten-Set, das die RFM-Scores enthält, enthält jetzt auch Name, Adresse und andere Informationen zu jedem Kunden.

Abbildung 8-10
Zusammengefügte Daten-Sets

Name	Typ	Spaltenformat	Dezimalstellen	Variablenlabel	Wertelabels
ID	Numerisch	5	0	Customer ID	Keine
Aktuellstes_Datum	Datum	10	0	Datum der letzt...	Keine
Transaktion_Häufigkeit	Numerisch	7	0	Anzahl an Tran...	Keine
Betrag	Numerisch	8	2	Betrag	Keine
Aktualitäts_Score	Numerisch	3	0	Aktualitäts-Score	Keine
Häufigkeits_Score	Numerisch	3	0	Häufigkeits-Score	Keine
Geldwert_Score	Numerisch	3	0	Geldwert-Score	Keine
RFM_Score	Numerisch	3	0	RFM-Score	Keine
Gender	Numerisch	4	0		{0, Female}...
AgeCategory	Numerisch	7	0	Age Category	{1, <25}...
Name	String	4	0		Keine
Address	String	14	0		Keine
City	String	11	0		Keine
State_Province	String	7	0	State/Province	Keine
PostalCode	String	1	0	Postal Code	Keine

Kapitel

9

Cluster-Analyse

Bei der Cluster-Analyse handelt es sich um eine explorative Prozedur zum Ermitteln von natürlichen Gruppierungen (Clustern) innerhalb Ihrer Daten. Damit können beispielsweise verschiedene Kundengruppen auf der Basis unterschiedlicher demographischer und Kaufverhaltensmerkmale ausgemacht werden.

Zum Beispiel möchte die Direktmarketing-Abteilung eines Unternehmens demografische Gruppierungen in ihrer Kundendatenbank identifizieren, um geeignete Strategien für ihre Marketingkampagnen zu ermitteln und neue Produktangebote zu entwickeln.

Diese Informationen finden Sie in der Datei *dmdata.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

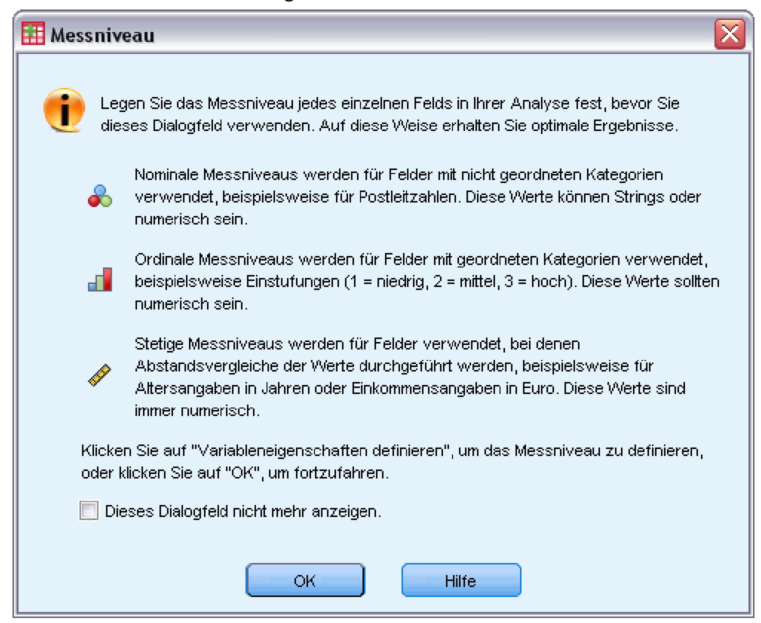
Durchführen der Analyse

▶ Zum Ausführen einer Cluster-Analyse wählen Sie die folgenden Menübefehle aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Meine Kontakte in Cluster segmentieren aus und klicken Sie auf Weiter.

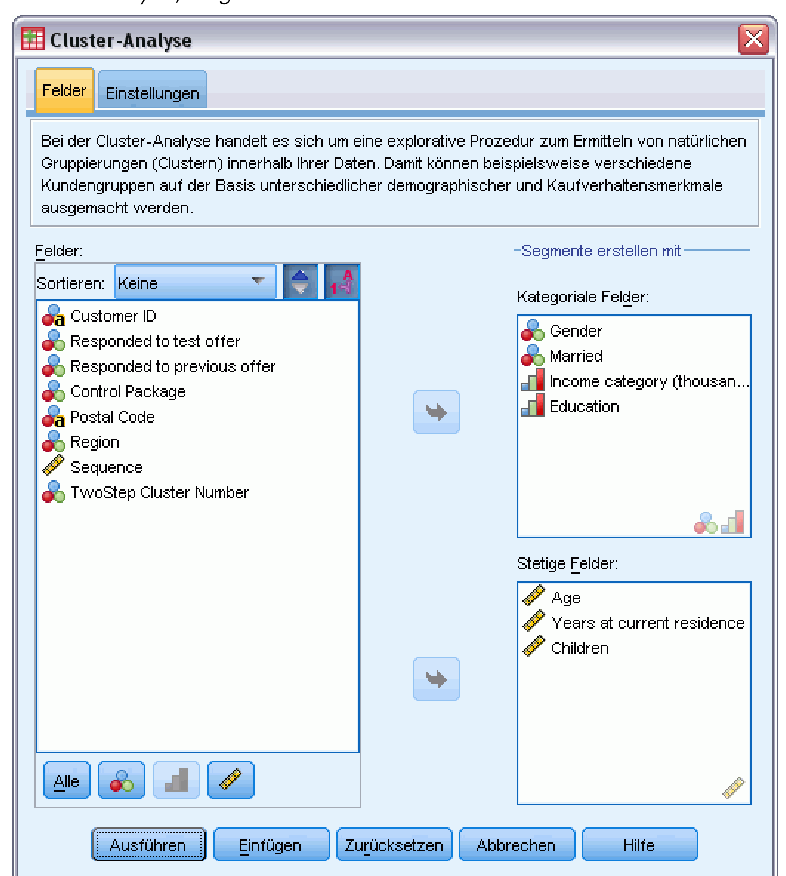
Abbildung 9-1
Messniveau-Warnmeldung



Die Messniveau-Warnmeldung wird angezeigt, weil das Messniveau die Berechnung der Ergebnisse für diese Prozedur beeinträchtigt. PASW Statistics weist neuen Daten-Sets anhand einiger einfacher Regeln automatisch ein Messniveau zu, wobei es sich hierbei nicht notwendigerweise um das beste Messniveau für alle Felder in Ihrem Daten-Set handeln muss. In diesem Beispiel-Daten-Set wurde allen Feldern das korrekte Messniveau zugewiesen.

▶ Klicken Sie zum Fortfahren auf OK.

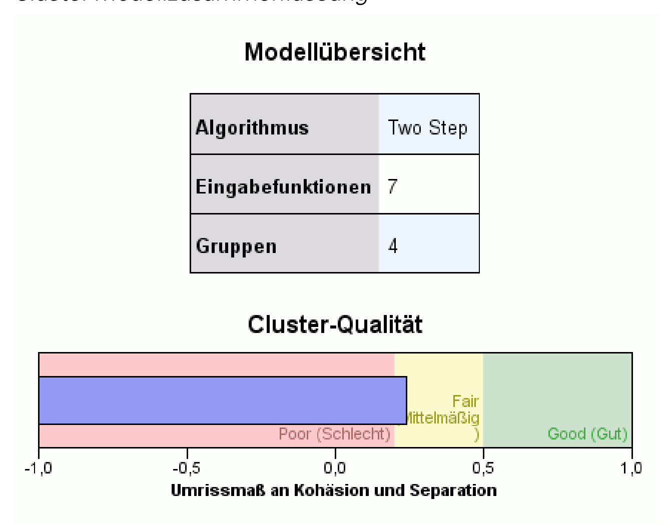
Abbildung 9-2
Cluster-Analyse, Registerkarte "Felder"



- Wählen Sie in der Symbolleiste "Felder" Messung aus der Dropdown-Liste "Sortieren" aus.
- Wählen Sie als kategoriale Felder *Geschlecht*, *Verheiratet*, *Einkommensklasse* und *Schulabschluss* aus.
- Wählen Sie als stetige Felder *Alter*, *Jahre an aktuellem Wohnort* und *Kinder* aus.
- Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Ausgabe

Abbildung 9-3
Cluster-Modellzusammenfassung

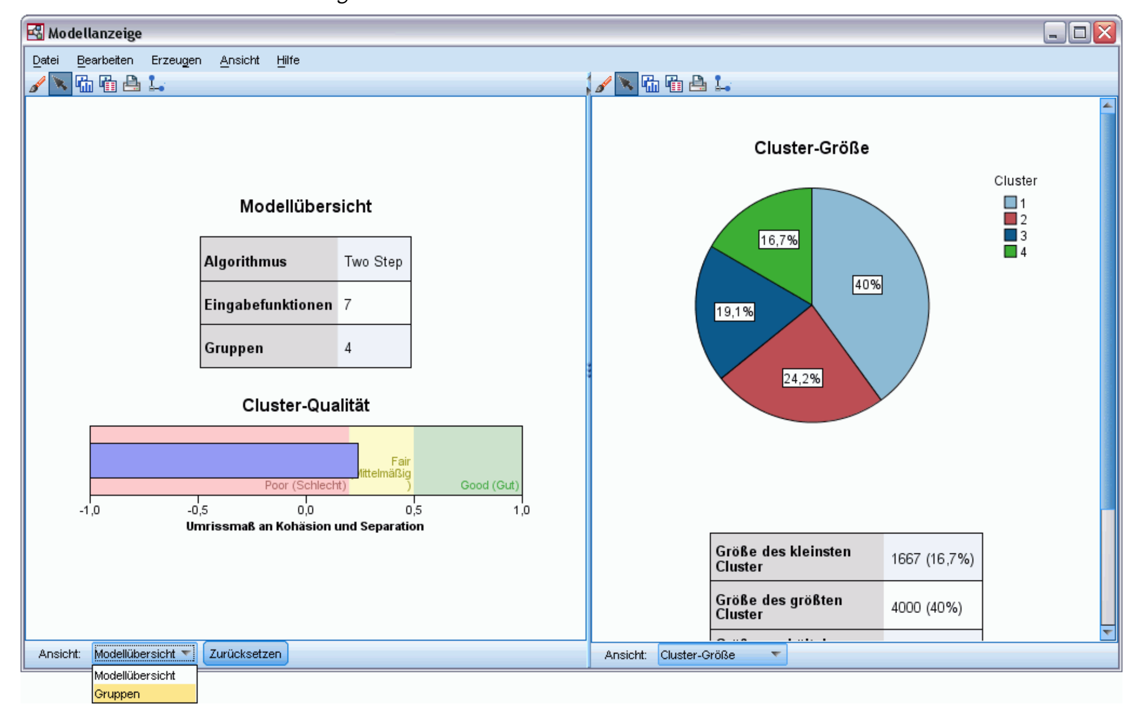


Die Ergebnisse werden in der Cluster-Modellanzeige angezeigt.

- Die Modellzusammenfassung zeigt, dass vier Cluster auf der Basis der sieben von Ihnen ausgewählten Eingabefunktionen (Eingabefelder) gefunden wurden.
- Das Diagramm zur Cluster-Qualität zeigt, dass die Gesamtqualität für das Modell im mittleren Bereich von “Fair” (Mittelmäßig) liegt.

▶ Doppelklicken Sie auf die Ausgabe der Cluster-Modellanzeige, um die Modellanzeige zu aktivieren.

Abbildung 9-4
Aktivierte Cluster-Modellanzeige



▶ Wählen Sie Cluster aus der Dropdown-Liste “Ansicht” im unteren Bereich des Fensters der Cluster-Modellanzeige aus.

Abbildung 9-5
Clusteransicht

Cluster	1	2	3	4
Beschriftung				
Beschreibung				
Größe	40,0% (4000)	24,2% (2424)	19,1% (1909)	16,7% (1667)
Funktionen	Age 50,30	Age 44,07	Age 39,05	Age 33,09
	Children 1,58	Children 1,29	Children 0,39	Children 0,12
	Gender Male (57,0%)	Gender Female (100,0%)	Gender Male (100,0%)	Gender Female (50,9%)
	Income category (thousands) 75+ (56,1%)	Income category (thousands) 50-74 (47,2%)	Income category (thousands) 75+ (34,8%)	Income category (thousands) <25 (100,0%)
	Married Yes (100,0%)	Married No (78,5%)	Married No (100,0%)	Married No (78,5%)
	Education Post-graduate (20,5%)	Education Post-graduate (20,5%)	Education College (21,1%)	Education Post-graduate (20,6%)
	Years at current residence 9,47	Years at current residence 9,51	Years at current residence 9,47	Years at current residence 9,42

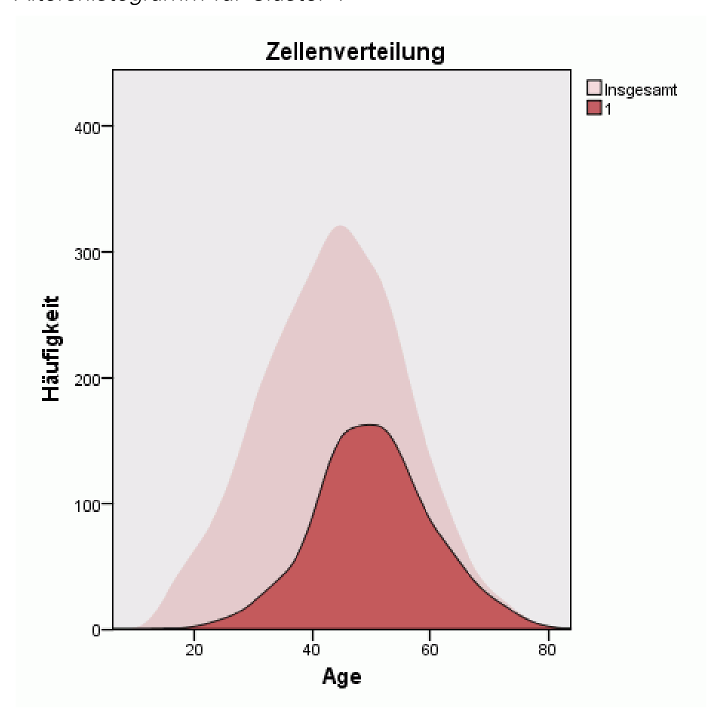
Die Clusteransicht enthält Informationen über die Attribute jedes Clusters.

- Bei stetigen (metrischen) Feldern wird der Mittelwert (Durchschnitt) angezeigt.
- Bei kategorialen Feldern (nominal, ordinal) wird der Modalwert angezeigt. Der Modalwert ist die Kategorie mit der größten Anzahl von Datensätzen. In diesem Beispiel entspricht jeder Datensatz einem Kunden.
- Standardmäßig werden Felder in der Reihenfolge ihrer Gesamtwichtigkeit für das Modell angezeigt. In diesem Beispiel hat *Alter* die größte Gesamtwichtigkeit. Sie können Felder auch nach Wichtigkeit innerhalb der Cluster oder in alphabetischer Reihenfolge sortieren.

Wenn Sie eine beliebige Zelle in der Clusteransicht auswählen, sehen Sie ein Diagramm, das die Werte dieses Felds für dieses Cluster zusammenfasst.

▶ Wählen Sie zum Beispiel die Zelle *Alter* für Cluster 1 aus.

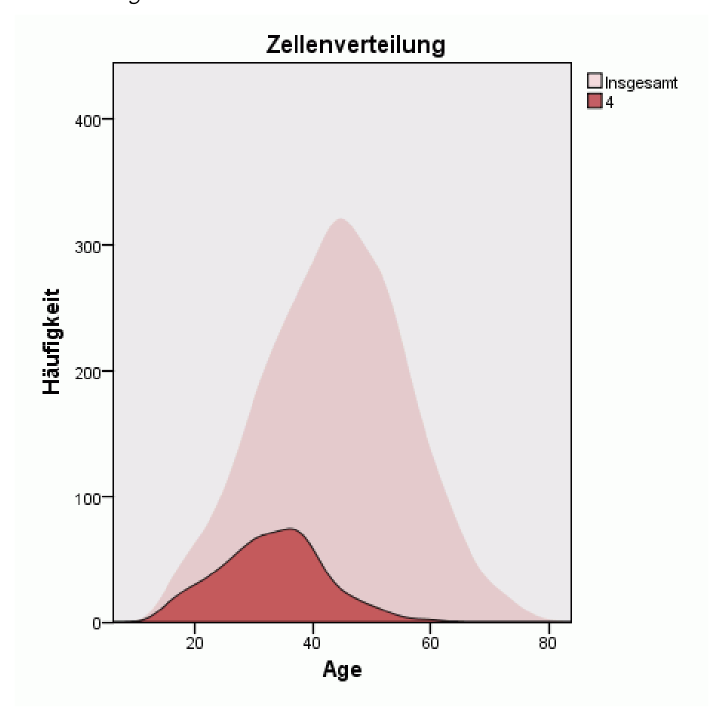
Abbildung 9-6
Altershistogramm für Cluster 1



Bei stetigen Feldern wird ein Histogramm angezeigt. Das Histogramm enthält sowohl die Verteilung von Werten innerhalb dieses Clusters als auch die Gesamtverteilung von Werten für das Feld. Das Histogramm zeigt, dass die Kunden in Cluster 1 tendenziell älter sind.

▶ Wählen Sie die Zelle *Alter* für Cluster 4 in der Clusteransicht aus.

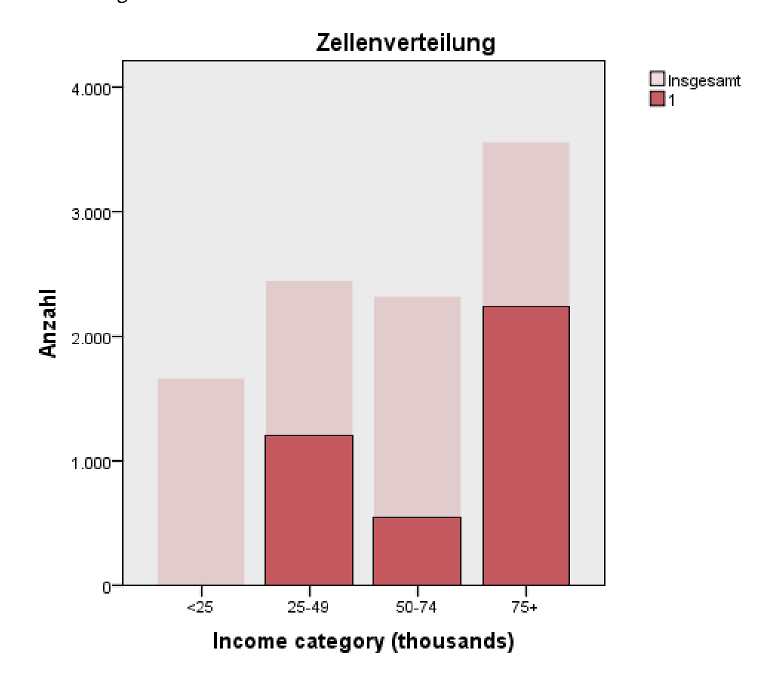
Abbildung 9-7
Altershistogramm für Cluster 4



Im Gegensatz zu Cluster 1 sind die Kunden in Cluster 4 tendenziell jünger als der Gesamtdurchschnitt.

- Wählen Sie die Zelle *Einkommensklasse* für Cluster 1 in der Clusteransicht aus.

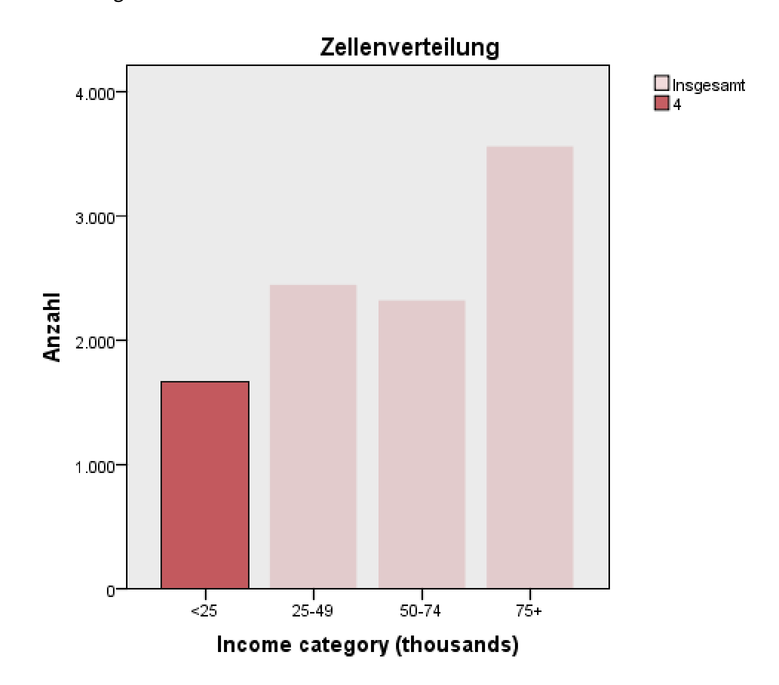
Abbildung 9-8
Balkendiagramm "Einkommensklasse" für Cluster 1



Bei kategorialen Feldern wird ein Balkendiagramm angezeigt. Das Bemerkenswerteste in dem Balkendiagramm "Einkommensklasse" für dieses Cluster ist, dass keinerlei Kunden in der niedrigsten Einkommensklasse vertreten sind.

► Wählen Sie die Zelle *Einkommensklasse* für Cluster 4 in der Clusteransicht aus.

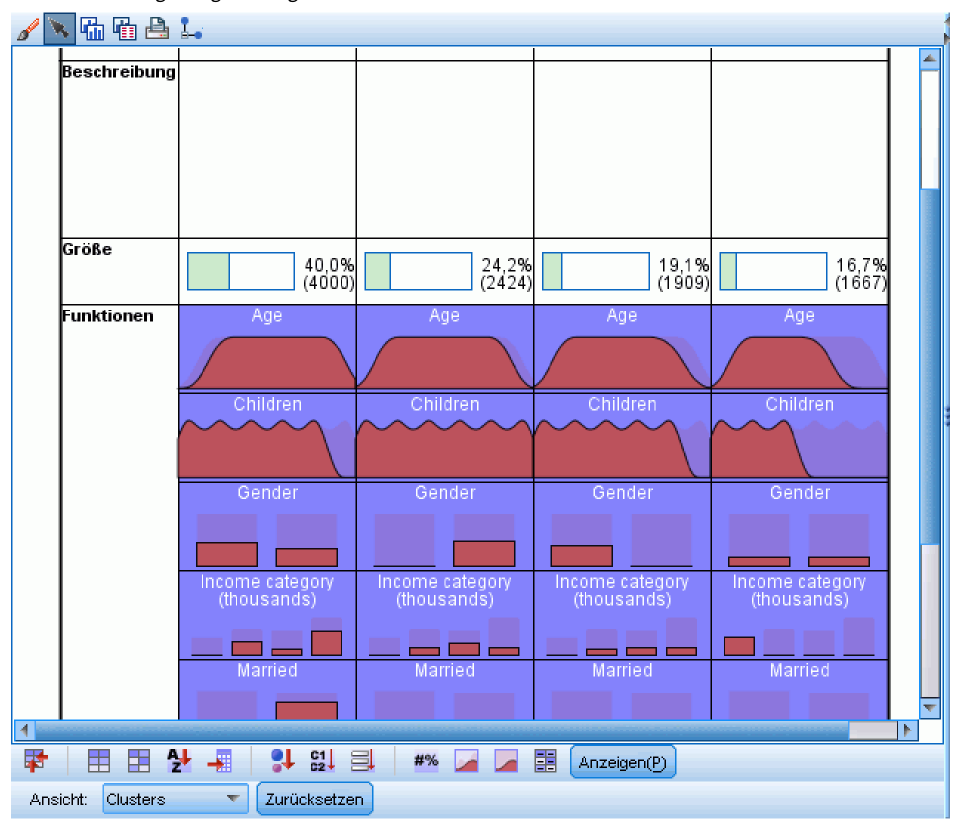
Abbildung 9-9
Balkendiagramm "Einkommensklasse" für Cluster 4



Im Gegensatz zu Cluster 1 sind alle Kunden in Cluster 4 in der niedrigsten Einkommensklasse vertreten.

Sie können die Clusteransicht auch so einstellen, dass Diagramme in der Zelle angezeigt werden. So lassen sich die Verteilungen von Werten zwischen Clustern schnell und einfach vergleichen, indem Sie die Symbolleiste im unteren Bereich des Fensters “Modellanzeige” zum Ändern der Ansicht verwenden.

Abbildung 9-10
Im Cluster angezeigte Diagramme



Wenn Sie einen genaueren Blick auf die Clusteransicht und die zusätzlichen in den Diagrammen für jede Zelle angezeigten Informationen werfen, erkennen Sie einige eindeutige Unterschiede zwischen den Clustern:

- Kunden in Cluster 1 sind tendenziell ältere, verheiratete Personen mit Kindern und höherem Einkommen.
- Kunden in Cluster 2 sind tendenziell ältere, allein erziehende Mütter mit durchschnittlichem Einkommen.
- Kunden in Cluster 3 sind tendenziell jüngere, allein stehende Männer ohne Kinder.
- Kunden in Cluster 4 sind tendenziell jüngere, allein stehende Frauen ohne Kinder und mit geringerem Einkommen.

Die Beschreibungszellen in der Clusteransicht sind Textfelder, die Sie bearbeiten können, um Beschreibungen jedes Clusters hinzuzufügen.

Abbildung 9-11
Clusteransicht mit Clusterbeschreibungen

Cluster	1	2	3	4
Beschriftung				
Beschreibung	Older, married, have children, higher income	Older single mothers, moderate income	Younger single men, no childern	Younger single women, no childern, low income
Größe	40,0% (4000)	24,2% (2424)	19,1% (1909)	16,7% (1667)
Funktionen	Age 50,30	Age 44,07	Age 39,05	Age 33,09
	Children 1,58	Children 1,29	Children 0,39	Children 0,12
	Gender Male (57,0%)	Gender Female (100,0%)	Gender Male (100,0%)	Gender Female (50,9%)
	Income category (thousands) 75+ (56,1%)	Income category (thousands) 50-74 (47,2%)	Income category (thousands) 75+ (34,8%)	Income category (thousands) <25 (100,0%)
	Married Yes (100,0%)	Married No (78,5%)	Married No (100,0%)	Married No (78,5%)
	Education Post-graduate (20,5%)	Education Post-graduate (20,5%)	Education College (21,1%)	Education Post-graduate (20,6%)
	Years at current residence 9,47	Years at current residence 9,51	Years at current residence 9,47	Years at current residence 9,42

Auswahl von Datensätzen auf der Basis von Clustern

Sie können Datensätze auf der Basis der Cluster-Zugehörigkeit auf zwei Arten auswählen:

- Erstellen Sie interaktiv eine Filterbedingung in der Cluster-Modellanzeige.
- Verwenden Sie die Werte des von der Prozedur erzeugten Clusterfelds, um Filter- oder Auswahlbedingungen zu bestimmen.

Erstellen eines Filters in der Cluster-Modellanzeige

So erstellen Sie eine Filterbedingung, die Datensätze aus bestimmten Clustern in der Cluster-Modellanzeige auswählt:

▶ Aktivieren Sie die Cluster-Modellanzeige durch Doppelklicken.

▶ Wählen Sie Cluster aus der Dropdown-Liste "Ansicht" im unteren Bereich des Fensters der Cluster-Modellanzeige aus.

▶ Klicken Sie im oberen Bereich der Clusteransicht auf die Clusternummer für das von Ihnen gewünschte Cluster. Wenn Sie mehrere Cluster auswählen möchten, klicken Sie bei gedrückter Strg-Taste auf jede zusätzliche von Ihnen gewünschte Clusternummer.

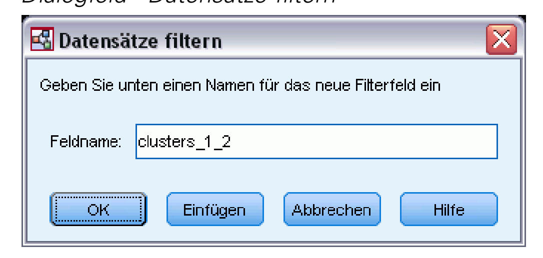
Abbildung 9-12
In der Clusteransicht ausgewählte Cluster

Cluster	1	2	3	4
Beschriftung				
Beschreibung	Older, married, have children, higher income	Older single mothers, moderate income	Younger single men, no childern	Younger single women, no childern, low income
Größe	40,0% (4000)	24,2% (2424)	19,1% (1909)	16,7% (1667)
Funktionen	Age 50,30	Age 44,07	Age 39,05	Age 33,09
	Children 1,58	Children 1,29	Children 0,39	Children 0,12
	Gender Male (57,0%)	Gender Female (100,0%)	Gender Male (100,0%)	Gender Female (50,9%)
	Income category (thousands) 75+ (56,1%)	Income category (thousands) 50-74 (47,2%)	Income category (thousands) 75+ (34,8%)	Income category (thousands) <25 (100,0%)
	Married Yes (100,0%)	Married No (78,5%)	Married No (100,0%)	Married No (78,5%)
	Education Post-graduate (20,5%)	Education Post-graduate (20,5%)	Education College (21,1%)	Education Post-graduate (20,6%)
	Years at current residence 9,47	Years at current residence 9,51	Years at current residence 9,47	Years at current residence 9,42

▶ Wählen Sie die folgenden Befehle aus den Menüs der Cluster-Modellanzeige aus:

Erzeugen
  Datensätze filtern

Abbildung 9-13
Dialogfeld “Datensätze filtern”



▶ Geben Sie einen Namen für das Filterfeld ein und klicken Sie auf OK. Die Namen müssen den Benennungsregeln von PASW Statistics entsprechen.

Abbildung 9-14
Gefilterte Datensätze im Daten-Editor

	ID	Married	Children	Region	ClusterGroup1	clusters_1_2
14	03623	No	0	West	3	.00
15	01353	No	0	West	3	.00
16	07055	No	0	West	3	.00
17	04455	No	0	West	2	1.00
18	07210	No	1	West	2	1.00
19	08054	No	0	West	4	.00
20	06937	No	0	West	4	.00
21	06512	No	0	West	4	.00
22	08315	No	0	West	4	.00
23	09676	No	3	West	2	1.00
24	09636	No	0	West	4	.00
25	08579	No	1	West	2	1.00
26	01480	No	1	West	2	1.00

Dadurch wird ein neues Feld im Daten-Set erzeugt und Datensätze werden anhand der Werte dieses Felds gefiltert.

- Datensätze mit dem Wert 1 für das Filterfeld werden in nachfolgende Analysen, Diagramme und Berichte aufgenommen.
- Datensätze mit dem Wert 0 für das Filterfeld werden ausgeschlossen.
- Ausgeschlossene Datensätze werden nicht aus dem Daten-Set entfernt, sondern mit einem Filterstatusindikator beibehalten, der als diagonaler Strich durch die Datensatznummer im Daten-Editor angezeigt wird.

Auswahl von Datensätzen auf der Basis von Clusterfeldwerten

Standardmäßig erstellt die Cluster-Analyse ein neues Feld, das die Clustergruppe für jeden Datensatz identifiziert. Der Standardname dieses Felds ist *ClusterGroupn*, wobei *n* eine Ganzzahl ist, die dem Feld einen eindeutigen Namen gibt.

Abbildung 9-15
Zum Daten-Set hinzugefügtes Clusterfeld

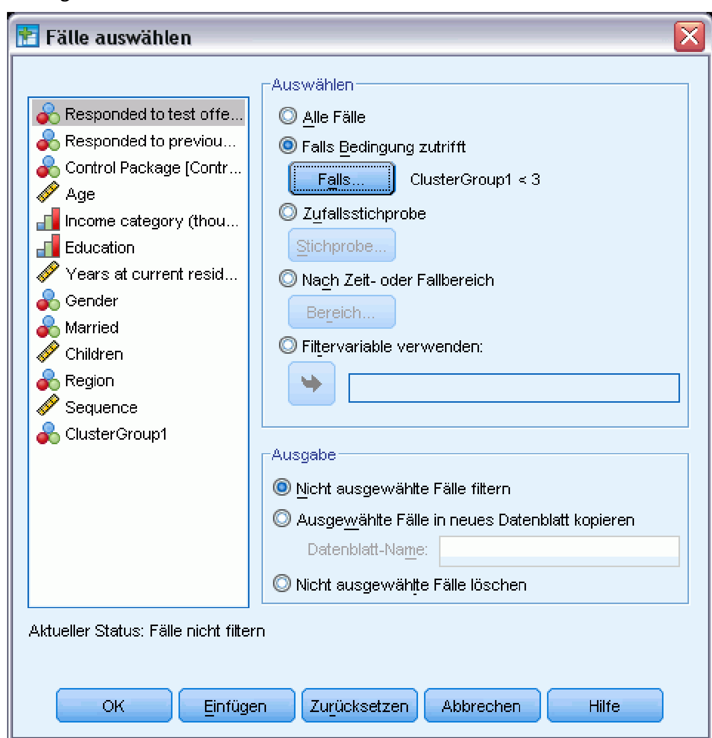
	ID	Gender	Married	Children	Region	ClusterGroup1
1	01359	Female	No	0	West	4
2	06262	Female	No	1	West	2
3	08031	Male	No	0	West	3
4	01971	Male	No	0	West	4
5	09689	Male	No	0	West	3
6	06108	Male	No	1	West	3
7	09853	Male	No	0	West	3
8	06802	Male	No	0	West	4
9	07597	Male	No	0	West	3
10	03692	Male	No	1	West	3
11	00071	Male	No	0	West	4
12	00769	Male	No	0	West	3

So verwenden Sie die Werte des Clusterfelds zur Auswahl von Datensätzen in bestimmten Clustern:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

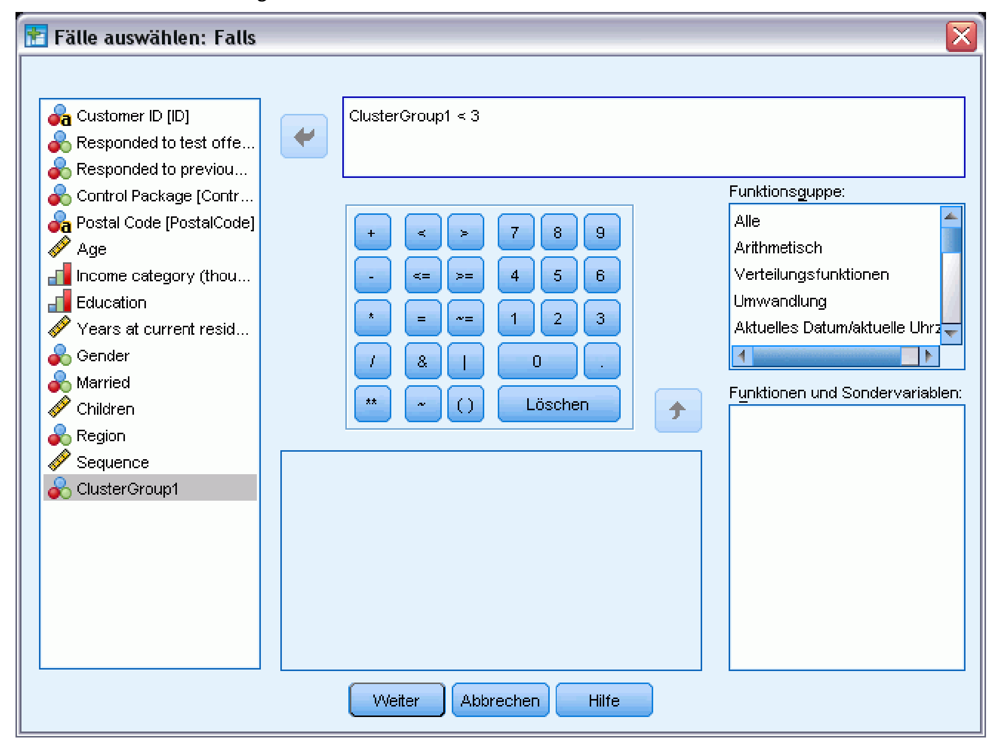
Daten
  Fälle auswählen

Abbildung 9-16
Dialogfeld "Fälle auswählen"



▶ Wählen Sie im Dialogfeld "Fälle auswählen" Falls Bedingung zutrifft und klicken Sie anschließend auf Falls.

Abbildung 9-17
Fälle auswählen: Dialogfeld “Falls”



- Geben Sie die Auswahlbedingung ein.

  Zum Beispiel werden mit ClusterGroup1 < 3 alle Datensätze in den Clustern 1 und 2 ausgewählt und Datensätze in den Clustern 3 und höher ausgeschlossen.

- Klicken Sie auf Weiter.

Im Dialogfeld “Fälle auswählen” gibt es mehrere Möglichkeiten, wie mit ausgewählten und nicht ausgewählten Datensätzen verfahren wird:

Nicht ausgewählte Fälle filtern. Hiermit wird ein neues Feld erstellt, das eine Filterbedingung angibt. Ausgeschlossene Datensätze werden nicht aus dem Daten-Set entfernt. sondern mit einem Filterstatusindikator beibehalten, der als diagonaler Strich durch die Datensatznummer im Daten-Editor angezeigt wird. Dies entspricht der interaktiven Auswahl von Clustern in der Cluster-Modellanzeige.

Kopieren von ausgewählten Fällen in ein neues Daten-Set. Hiermit wird ein neues Daten-Set in der aktuellen Sitzung erstellt, das nur die Datensätze enthält, die die Filterbedingung erfüllen. Das ursprüngliche Daten-Set bleibt davon unberührt.

Nicht ausgewählte Fälle löschen. Nicht ausgewählte Datensätze werden aus dem Daten-Set gelöscht. Gelöschte Datensätze können nur wiederhergestellt werden, indem Sie die Datei ohne Speichern der Änderungen schließen und sie dann erneut öffnen. Wenn Sie die Änderungen in der Datendatei speichern, werden die Fälle dauerhaft gelöscht.

Das Dialogfeld “Fälle auswählen” verfügt über eine Option zur Verwendung einer bestehenden Variable als Filtervariable (Variablenfeld). Wenn Sie interaktiv eine Filterbedingung in der Cluster-Modellanzeige erstellen und das erzeugte Filterfeld im Daten-Set speichern, können Sie dieses Feld verwenden, um Datensätze in Folgesitzungen zu filtern.

Zusammenfassung

Bei der Cluster-Analyse handelt es sich um eine nützliche explorative Prozedur zum Ermitteln von natürlichen Gruppierungen (Clustern) innerhalb Ihrer Daten. Sie können mithilfe der Informationen aus diesen Clustern geeignete Strategien für Ihre Marketingkampagnen ermitteln und neue Produktangebote entwickeln. Sie können Datensätze anhand der Cluster-Zugehörigkeit zur weiteren Analyse oder für ausgerichtete Marketingkampagnen auswählen.

Kapitel

10

Profile über potenzielle Kunden

Bei Profilen über potenzielle Kunden werden Ergebnisse aus einer früheren Kampagne oder einer Testkampagne verwendet, um beschreibende Profile zu erstellen. Diese Profile können bei zukünftigen Kampagnen für das Targeting bestimmter Gruppen von Kontakten verwendet werden. Zum Beispiel möchte die Marketing-Abteilung eines Unternehmens anhand der Ergebnisse einer Testsendung auf Basis von demografischen Informationen Profile der Typen von Personen erstellen, bei denen die Wahrscheinlichkeit einer Antwort auf ein bestimmtes Angebot am höchsten ist. Anhand dieser Ergebnisse können sie dann die Arten der Verteilerlisten ermitteln, die sie für ähnliche Angebote verwenden sollten.

Beispielsweise verschickt die Direktmarketing-Abteilung eines Unternehmens eine Testsendung an ca. 20 % ihrer gesamten Kundendatenbank. Die Ergebnisse dieser Testsendung werden in einer Datendatei aufgezeichnet, die außerdem demografische Merkmale eines jeden Kunden enthält, zum Beispiel Alter, Geschlecht, Familienstand und geografische Region. Die Ergebnisse werden auf einfache Weise mit Ja/Nein aufgezeichnet, um zu erfahren, welche Kunden in der Testsendung geantwortet (einen Kauf abgeschlossen) haben und welche nicht.

Diese Informationen finden Sie in der Datei *dmdata.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

Erläuterung der Daten

Das Responsefeld sollte kategorial sein, wobei ein Wert alle positiven Reaktionen darstellen sollte. Es wird davon ausgegangen, dass alle anderen nicht fehlenden Werte eine negative Antwort anzeigen. Falls das Responsefeld einen stetigen (metrischen) Wert enthält, beispielsweise die Anzahl oder den Geldwert der Käufe, müssen Sie ein neues Feld erstellen, das allen von Null abweichenden Responsewerten eine einzelne positive Antwort zuweist.Für weitere Informationen siehe Erstellen eines kategorialen Responsefelds in Kapitel 4 auf S. 24.

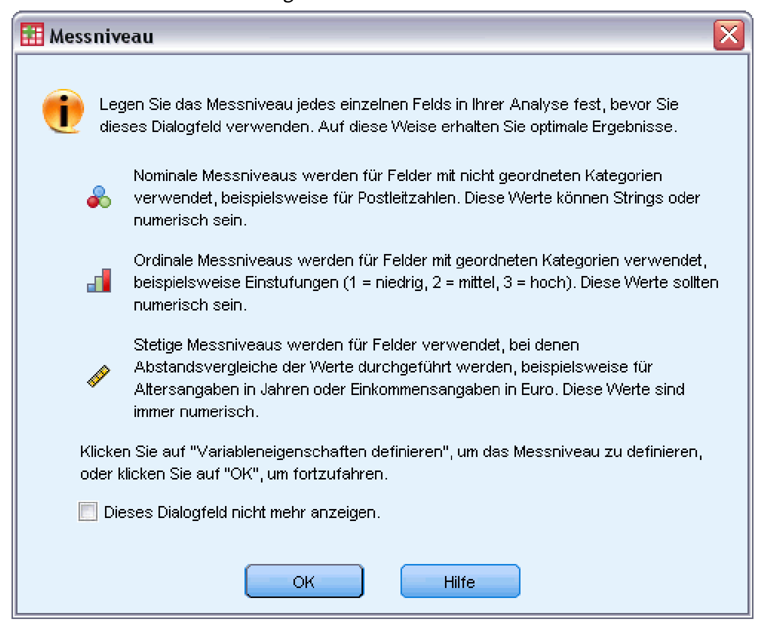
Durchführen der Analyse

▶ Um eine Analyse von Profilen über potenzielle Kunden auszuführen, wählen Sie in den Menüs folgende Optionen aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Profile für die Kontakte erstellen, die auf ein Angebot reagiert haben aus und klicken Sie auf Weiter.

Abbildung 10-1
Messniveau-Warnmeldung



Die Messniveau-Warnmeldung wird angezeigt, weil das Messniveau die Berechnung der Ergebnisse für diese Prozedur beeinträchtigt. PASW Statistics weist neuen Daten-Sets anhand einiger einfacher Regeln automatisch ein Messniveau zu, wobei es sich hierbei nicht notwendigerweise um das beste Messniveau für alle Felder in Ihrem Daten-Set handeln muss. In diesem Beispiel-Daten-Set wurde allen Feldern das korrekte Messniveau zugewiesen.

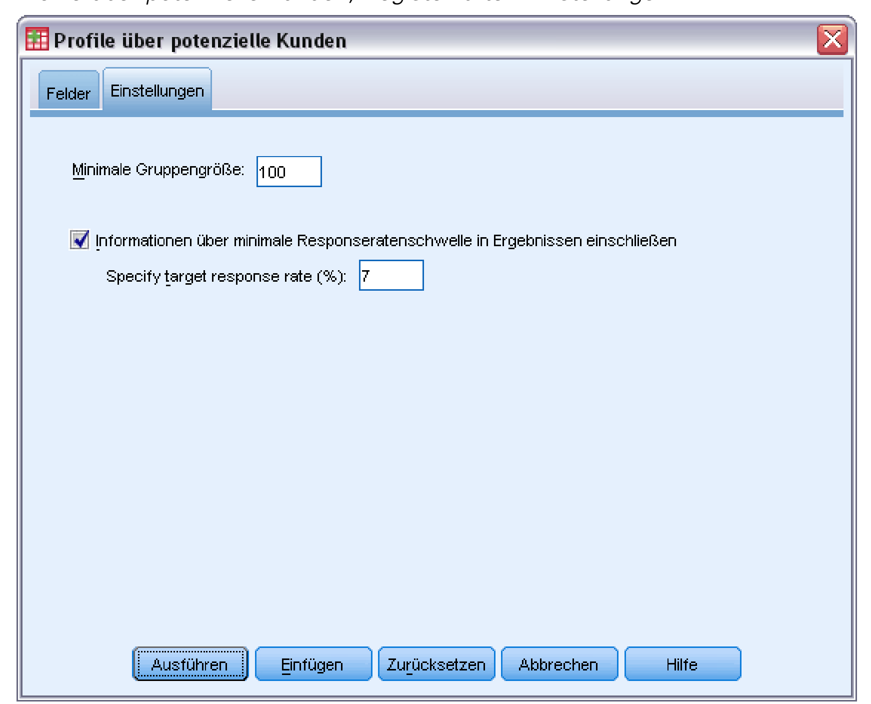
▶ Klicken Sie zum Fortfahren auf OK.

Abbildung 10-2
Profile über potenzielle Kunden, Registerkarte "Felder"



- Wählen Sie bei "Responsefeld" *Auf Testangebot geantwortet* aus.
- Wählen Sie bei "Wert für positive Antworten" *Ja* aus der Dropdown-Liste aus. Im Textfeld wird der Wert 1 angezeigt, da es sich bei "Ja" eigentlich um ein Wertelabel handelt, das zum aufgezeichneten Wert 1 gehört. (Wenn für den Wert für positive Antworten kein Wertelabel definiert wurde, können Sie den Wert einfach in das Textfeld eingeben.)
- Wählen Sie bei "Profile erstellen mit" *Alter*, *Einkommensklasse*, *Schulbildung*, *Jahre an aktuellem Wohnort*, *Geschlecht*, *Verheiratet*, *Region* und *Kinder* aus.
- Klicken Sie auf die Registerkarte Einstellungen.

Abbildung 10-3
Profile über potenzielle Kunden, Registerkarte "Einstellungen"



- Aktivieren Sie "Informationen über minimale Responseratenschwelle in Ergebnissen einschließen".
- Geben Sie als Ziel-Responserate den Wert 7 ein.
- Klicken Sie dann auf Ausführen, um die Prozedur auszuführen.

Ausgabe

Abbildung 10-4
Tabelle für die Responserate

Responserate

Nummer	Profil			
	Beschreibung	Gruppengröße	Responserate	Kumulierte Responserate
1	Region = "West","South","East" Gender = "Female" Married = "No"	379	9.2%	9.2%
2	Region = "West","South","East" Gender = "Female" Married = "Yes"	299	5.0%	7.4%
3	Region = "West","South","East" Gender = "Male"	722	4.7%	6.0%
4	Region = "North"	517	2.5%	5.1%

Grün: Erfüllt die Ziel-Responserate.
Rot: Erfüllt die Ziel-Responserate nicht.

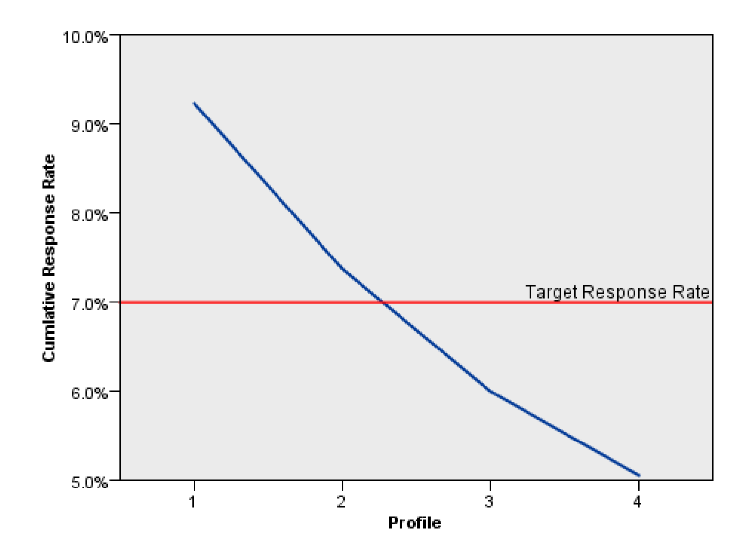
In der Tabelle für die Responserate werden Informationen für jede durch die Prozedur identifizierte Profilgruppe angezeigt.

- Profile werden in absteigender Reihenfolge der Responserate angezeigt.
- Die Responserate ist der Prozentsatz von Kunden, die positiv reagiert (einen Kauf abgeschlossen) haben.
- Die kumulative Responserate ist die kombinierte Responserate für die aktuelle und alle vorherigen Profilgruppen. Da die Profile in absteigender Reihenfolge der Responserate angezeigt werden, handelt es sich bei der kumulativen Responserate um die kombinierte Responserate für die aktuelle Profilgruppe plus aller Profilgruppen mit einer höheren Responserate.
- Die Profilbeschreibung enthält nur die Merkmale für jene Felder, die einen signifikanten Beitrag zum Modell leisten. In diesem Beispiel sind Region, Geschlecht und Familienstand im Modell enthalten. Die restlichen Felder – "Alter", "Einkommen", "Schulbildung" und "Jahre an aktuellem Wohnort" – sind nicht enthalten, da sie keinen signifikanten Beitrag zum Modell geleistet haben.
- Der grüne Bereich der Tabelle entspricht den Profilen mit einer kumulativen Responserate größer oder gleich der angegebenen Ziel-Responserate, in diesem Beispiel 7 %.
- Der rote Bereich der Tabelle entspricht den Profilen mit einer kumulativen Responserate unter der angegebenen Ziel-Responserate.
- Die kumulative Responserate in der letzten Zeile der Tabelle ist die gesamte oder durchschnittliche Responserate für alle in die Testsendung aufgenommenen Kunden, da es sich dabei um die Responserate für alle Profilgruppen handelt.

Die in der Tabelle angezeigten Ergebnisse lassen darauf schließen, dass Sie bei einer weiblichen Zielgruppe im Westen, Süden und Osten eine Responserate erzielen sollten, die leicht über der Ziel-Responserate liegt.

Beachten Sie jedoch, dass es in diesen Regionen eine deutliche Abweichung zwischen der Responserate bei unverheirateten Frauen (9,2 %) und verheirateten Frauen (5,0 %) gibt. Obwohl die kumulative Responserate bei beiden Gruppen über der Ziel-Responserate liegt, ist die Responserate bei der letzten Gruppe allein tatasächlich niedriger als die Ziel-Responserate, was darauf schließen lässt, dass Sie andere Merkmale suchen sollten, um das Modell zu verbessern.

Abbildung 10-5
Diagramm mit kumulativer Responserate



Das Diagramm mit kumulativer Responserate ist im Wesentlichen eine visuelle Darstellung der in der Tabelle angezeigten Responseraten. Da die Profile in absteigender Reihenfolge der Responserate angezeigt werden, bewegt sich die Linie für die kumulative Responserate mit jedem weiteren Profil stets nach unten. Genau wie in der Tabelle zeigt sich auch im Diagramm, dass die kumulative Responserate unter die Ziel-Responserate zwischen Profilgruppe 2 und Profilgruppe 3 fällt.

Zusammenfassung

Bei dieser speziellen Testsendung wurden vier Profilgruppen identifiziert und die Ergebnisse zeigen, dass es sich bei den einzigen signifikanten demografischen Merkmalen, die damit in Zusammenhang stehen, ob eine Person auf ein Angebot reagiert hat oder nicht, um “Geschlecht”, “Region” und “Familienstand” handelt. Die Gruppe mit der höchsten Responserate besteht aus unverheirateten Frauen, die im Süden, Osten und Westen leben. Danach nehmen die Responseraten rapide ab, obwohl die Aufnahme von verheirateten Frauen in denselben Regionen dennoch zu einer kumulativen Responserate führt, die über der Ziel-Responserate liegt.

Kapitel



Responseraten nach Postleitzahlen

Bei dieser Technik werden Ergebnisse aus einer früheren Kampagne verwendet, um Responseraten nach Postleitzahlen zu berechnen. Diese Raten können bei zukünftigen Kampagnen für das Targeting bestimmter Postleitzahlbereiche verwendet werden.

Beispielsweise erzeugt die Marketing-Abteilung eines Unternehmens anhand der Ergebnisse einer früheren Postsendungs-Kampagne Responseraten nach Postleitzahlen. Auf Basis verschiedener Kriterien wie der minimalen akzeptablen Responserate und/oder der maximalen Anzahl von Kontakten, die in die Postsendungs-Kampagne eingeschlossen werden sollen, können daraufhin bestimmte Postleitzahlbereiche für die Kampagne bestimmt werden.

Diese Informationen finden Sie in der Datei *dmdata.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

Erläuterung der Daten

Das Responsefeld sollte kategorial sein, wobei ein Wert alle positiven Reaktionen darstellen sollte. Es wird davon ausgegangen, dass alle anderen nicht fehlenden Werte eine negative Antwort anzeigen. Falls das Responsefeld einen stetigen (metrischen) Wert enthält, beispielsweise die Anzahl oder den Geldwert der Käufe, müssen Sie ein neues Feld erstellen, das allen von Null abweichenden Responsewerten eine einzelne positive Antwort zuweist. Für weitere Informationen siehe Erstellen eines kategorialen Responsefelds in Kapitel 5 auf S. 33.

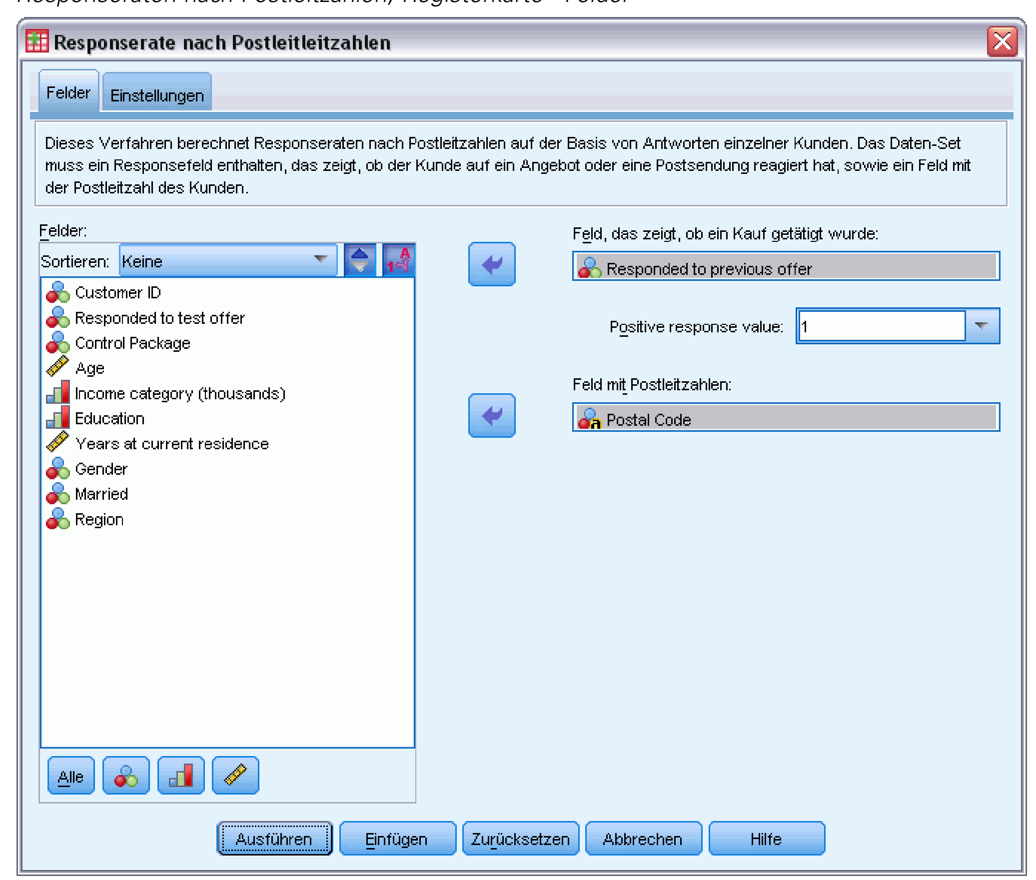
Durchführen der Analyse

▶ Um Responseraten nach Postleitzahlen zu berechnen, wählen Sie in den Menüs folgende Optionen aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

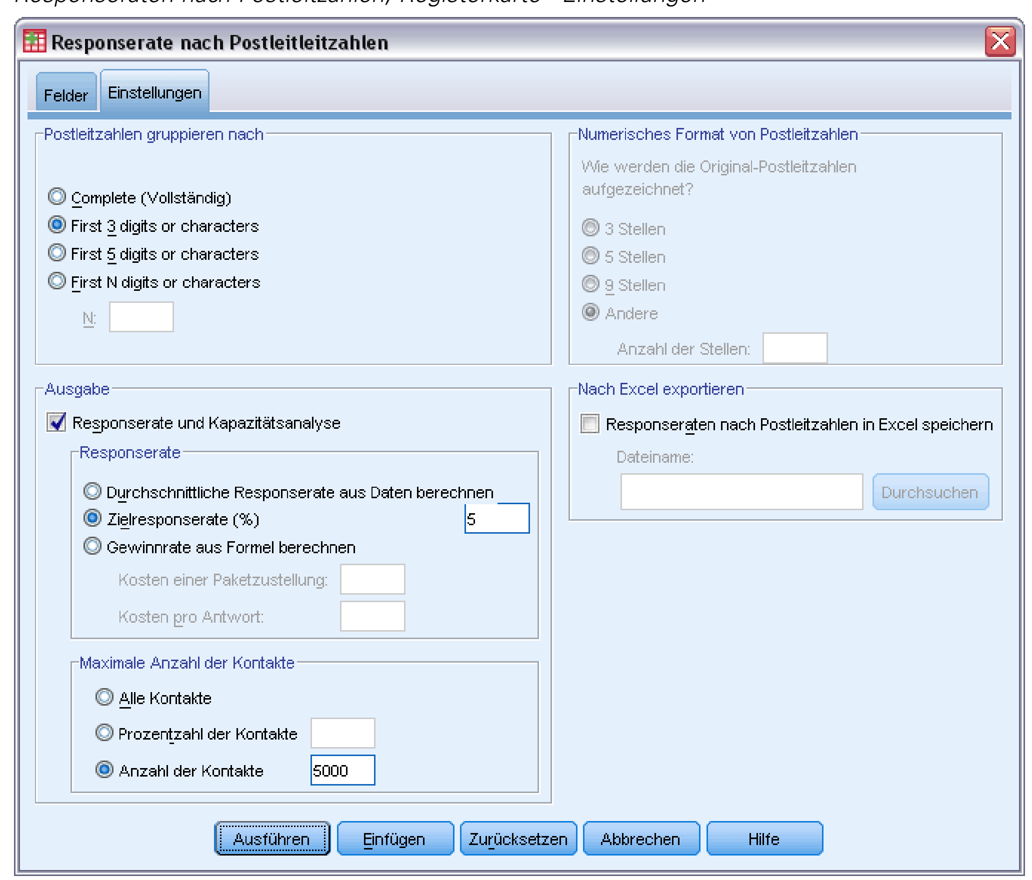
▶ Wählen Sie Postleitzahlbereiche mit den meisten Antworten identifizieren aus und klicken Sie auf Weiter.

Abbildung 11-1
Responseraten nach Postleitzahlen, Registerkarte "Felder"



- Wählen Sie bei "Responsefeld" *Auf vorheriges Angebot geantwortet* aus.
- Wählen Sie bei "Wert für positive Antworten" *Ja* aus der Dropdown-Liste aus. Im Textfeld wird der Wert 1 angezeigt, da es sich bei "Ja" eigentlich um ein Wertelabel handelt, das zum aufgezeichneten Wert 1 gehört. (Wenn für den Wert für positive Antworten kein Wertelabel definiert wurde, können Sie den Wert einfach in das Textfeld eingeben.)
- Wählen Sie bei "Postleitzahlfeld" *Postleitzahl* aus.
- Klicken Sie auf die Registerkarte Einstellungen.

Abbildung 11-2
Responseraten nach Postleitzahlen, Registerkarte "Einstellungen"



- Wählen Sie in der Gruppe "Postleitzahlen gruppieren nach" Die ersten 3 Stellen oder Zeichen aus. Dadurch werden kombinierte Responseraten für alle Kontakte berechnet, deren Postleitzahlen mit den gleichen drei Stellen oder Zeichen beginnen. Beispielsweise stellen die ersten drei Stellen einer US-amerikanischen Postleitzahl eine gemeinsame geografische Region dar, die größer ist als die durch die vollständige fünfstellige Postleitzahl definierte geografische Region.

- Aktivieren Sie in der Gruppe "Ausgabe" "Responserate und Kapazitätsanalyse".

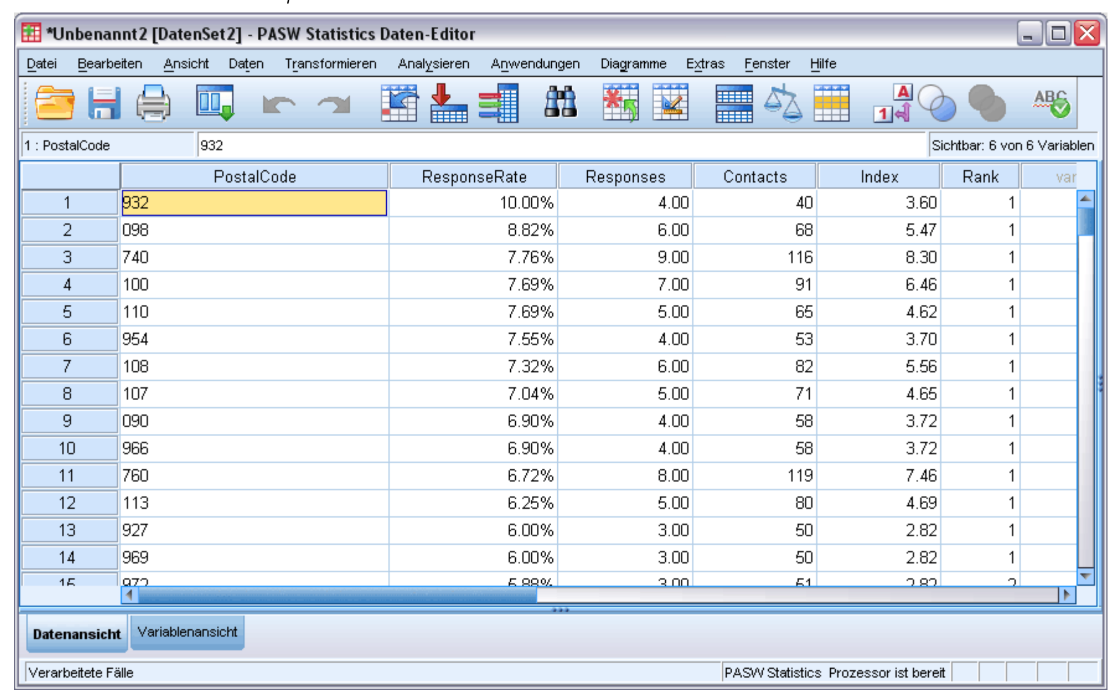
- Wählen Sie "Ziel-Responserate" aus und geben Sie den Wert 5 ein.

- Wählen Sie "Anzahl der Kontakte" aus und geben Sie den Wert 5000 ein.

- Klicken Sie dann auf Ausführen, um die Prozedur auszuführen.

Ausgabe

Abbildung 11-3
Neues Daten-Set mit Responseraten nach Postleitzahlen



	PostalCode	ResponseRate	Responses	Contacts	Index	Rank
1	932	10.00%	4.00	40	3.60	1
2	098	8.82%	6.00	68	5.47	1
3	740	7.76%	9.00	116	8.30	1
4	100	7.69%	7.00	91	6.46	1
5	110	7.69%	5.00	65	4.62	1
6	954	7.55%	4.00	53	3.70	1
7	108	7.32%	6.00	82	5.56	1
8	107	7.04%	5.00	71	4.65	1
9	090	6.90%	4.00	58	3.72	1
10	966	6.90%	4.00	58	3.72	1
11	760	6.72%	8.00	119	7.46	1
12	113	6.25%	5.00	80	4.69	1
13	927	6.00%	3.00	50	2.82	1
14	969	6.00%	3.00	50	2.82	1

Es wird automatisch ein neues Daten-Set erstellt. Dieses Daten-Set enthält einen einzelnen Datensatz (Zeile) für jede Postleitzahl. In diesem Beispiel enthält jede Zeile Auswertungsinformationen für alle Postleitzahlen, die mit den gleichen drei Stellen oder Zeichen beginnen.

Zusätzlich zu dem Feld mit der Postleitzahl enthält das neue Daten-Set die folgenden Felder:

- **Responserate.** Der Prozentsatz der positiven Antworten in jeder Postleitzahl-Gruppe. Datensätze werden automatisch in absteigender Reihenfolge der Responseraten sortiert, d. h., Postleitzahlen mit der höchsten Responserate erscheinen am Anfang des Daten-Sets.
- **Antworten.** Der Anzahl der positiven Antworten in jeder Postleitzahl-Gruppe.
- **Kontakte.** Die Gesamtanzahl von Kontakten in jedem Postleitzahlbereich, die einen nicht fehlenden Wert für das Responsefeld enthalten.
- **Index.** Die “gewichtete” Antwort auf Basis der Formel *N x P x (1-P)*, wobei *N* die Anzahl von Kontakten und *P* die als Anteil ausgedrückte Responserate ist. Bei zwei Postleitzahlen mit derselben Responserate weist diese Formel der Postleitzahl mit der höheren Anzahl an Kontakten einen höheren Indexwert zu.
- **Rang.** Dezil-Rang (oberste 10 %, oberste 20 % usw.) der kumulativen Postleitzahl-Responseraten in absteigender Reihenfolge.

Da auf der Registerkarte “Einstellungen” des Dialogfelds “Responseraten nach Postleitzahlen” die Option “Responserate und Kapazitätsanalyse” ausgewählt wurde, werden eine Auswertungstabelle und ein Auswertungsdiagramm für die Responserate im Viewer angezeigt.

Abbildung 11-4
Tabelle für die Responserate

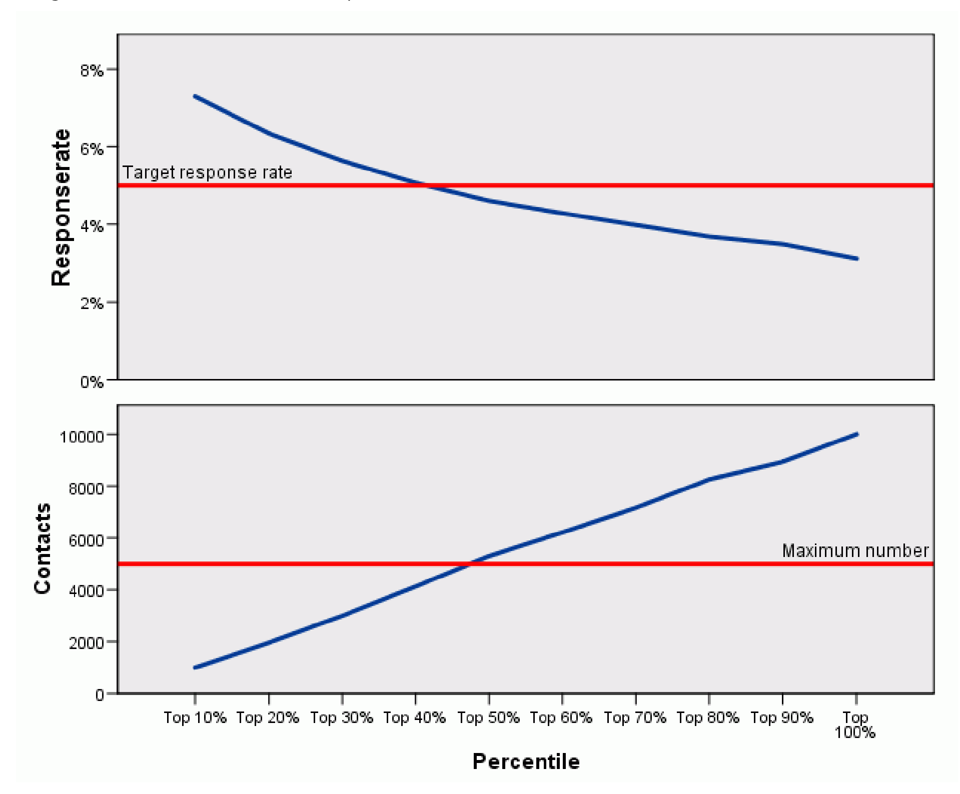
Percentile	Responserate	Contacts	Cumulative Response Rate	Total Contacts
Top 10%	7.3	1001	7.3	1001
Top 20%	5.3	956	6.3	1957
Top 30%	4.3	1042	5.6	2999
Top 40%	3.5	1127	5.1	4126
Top 50%	3.0	1173	4.6	5299
Top 60%	2.4	914	4.3	6213
Top 70%	2.0	948	4.0	7161
Top 80%	1.7	1095	3.7	8256
Top 90%	1.2	680	3.5	8936
Top 100%	.0	1064	3.1	10000

Grün: Erfüllt die Ziel-Responserate.
Rot: Erfüllt die Ziel-Responserate nicht.

In der Tabelle werden die Ergebnisse nach Dezil-Rang in absteigender Reihenfolge (die besten 10 %, die besten 20 % etc.) zusammengefasst.

- Die kumulative Responserate ist der kombinierte Prozentsatz der positiven Antworten in der aktuellen und in allen vorherigen Zeilen. Da die Ergebnisse in absteigender Reihenfolge der Responseraten angezeigt werden, handelt es sich hierbei folglich um die kombinierte Responserate für das aktuelle Dezil und alle Dezile mit einer höheren Responserate.
- Die Tabelle wird auf Basis der von Ihnen eingegebenen Werte für “Ziel-Responserate” und “Maximale Anzahl von Kontakten” farbkodiert. Zeilen mit einer kumulativen Responserate größer oder gleich 5 % und maximal 5.000 kumulativen Kontakten werden grün markiert. Die Farbkodierung richtet sich danach, welcher Schwellenwert zuerst erreicht wird. In diesem Beispiel werden beide Schwellenwerte im selben Dezil erreicht.

Abbildung 11-5
Diagramm mit kumulativer Responserate



Das Diagramm der kumulativen Responserate und der kumulativen Anzahl an Kontakten in jedem Dezil ist eine visuelle Darstellung der gleichen Informationen, die auch in der Tabelle für die Responserate angezeigt werden. Der Schwellenwert für die minimale kumulative Responserate und die maximale kumulative Anzahl von Kontakten liegt in etwa zwischen dem 40. und dem 50. Perzentil.

- Da in dem Diagramm kumulative Responseraten in absteigender Reihenfolge des Dezil-Rangs der Responserate angezeigt werden, geht die Linie der kumulativen Responserate stets mit jedem weiteren Dezil nach unten.
- Da die Linie der Anzahl von Kontakten die kumulative Anzahl von Kontakten darstellt, geht sie stets nach oben.

Anhand der Informationen in der Tabelle und dem Diagramm sehen Sie, dass Sie sich auf die Postleitzahlen in den ersten vier Dezilen konzentrieren sollten, wenn Sie eine Responserate von mindestens 5 % erreichen, aber nicht mehr als 5.000 Kontakte in die Kampagne aufnehmen

möchten. Da der Dezil-Rang im neuen Daten-Set enthalten ist, können Sie die Postleitzahlen, die die erforderlichen ersten 40 % erreichen, leicht identifizieren.

Abbildung 11-6
Neues Datenblatt

*Unbenannt9 [DatenSet1] - PASW Statistics Daten-Editor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Anwendungen Diagramme Extras Fenster Hilfe

1 : PostalCode | 932 | Sichtbar: 6 von 6 Variablen

	PostalCode	ResponseRate	Responses	Contacts	Index	Rank
48	120	3.57%	3.00	84	2.89	Top 40%
49	965	3.57%	2.00	56	1.93	Top 40%
50	618	3.54%	4.00	113	3.86	Top 40%
51	603	3.53%	3.00	85	2.89	Top 40%
52	757	3.48%	4.00	115	3.86	Top 40%
53	948	3.39%	2.00	59	1.93	Top 40%
54	103	3.33%	3.00	90	2.90	Top 40%
55	608	3.33%	3.00	90	2.90	Top 40%
56	612	3.28%	4.00	122	3.87	Top 50%
57	762	3.23%	1.00	31	.97	Top 50%
58	933	3.23%	2.00	62	1.94	Top 50%

Datenansicht | Variablenansicht

Anmerkung: Der *Rang* wird als ganzzahliger Wert zwischen 1 und 10 aufgezeichnet. Das Feld verfügt über definierte Wertelabels, wobei der Wert 1 den ersten 10 %, der Wert 2 den ersten 20 % usw. entspricht. Je nach Ihren Anzeigeeinstellungen sehen Sie entweder die tatsächlichen Rangwerte oder die Wertelabels in der Datenansicht des Daten-Editors.

Zusammenfassung

Bei der Prozedur “Responseraten nach Postleitzahlen” werden Ergebnisse aus einer früheren Kampagne verwendet, um Responseraten nach Postleitzahlen zu berechnen. Diese Raten können bei zukünftigen Kampagnen für das Targeting bestimmter Postleitzahlbereiche verwendet werden. Bei der Prozedur wird ein neues Daten-Set erstellt, das Responseraten für jede Postleitzahl enthält. Anhand der Informationen in der Tabelle und dem Diagramm für die Responserate sowie der Informationen des Dezil-Rangs im neuen Daten-Set können Sie all jene Postleitzahlen identifizieren, die die angegebene minimale kumulative Responserate und/oder die kumulative maximale Anzahl von Kontakten erreichen.

Kapitel

12

Kaufneigung

Für die Kaufneigung werden Ergebnisse einer Testsendung oder einer früheren Kampagne verwendet, um Neigungsbewertungen zu erstellen. Die Bewertungen zeigen anhand von zahlreichen ausgewählten Merkmalen an, bei welchen Kontakten die Wahrscheinlichkeit einer Antwort am höchsten ist. Bei dieser Technik wird die binäre logistische Regression verwendet, um die Neigungsbewertungen zu erstellen.

So verwendet zum Beispiel die Direktmarketing-Abteilung eines Unternehmens die Ergebnisse einer Testsendung, um den übrigen Kontakten in ihrer Datenbank Neigungsbewertungen zuzuweisen, wobei verschiedene demografische Merkmale eingesetzt werden, um Kontakte zu ermitteln, bei denen die Wahrscheinlichkeit einer Antwort und eines Kaufs am größten ist.

Diese Informationen finden Sie in der Datei *dmdata.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

Erläuterung der Daten

Das Responsefeld sollte kategorial sein, wobei ein Wert alle positiven Reaktionen darstellen sollte. Es wird davon ausgegangen, dass alle anderen nicht fehlenden Werte eine negative Antwort anzeigen. Falls das Responsefeld einen stetigen (metrischen) Wert enthält, beispielsweise die Anzahl oder den Geldwert der Käufe, müssen Sie ein neues Feld erstellen, das allen von Null abweichenden Responsewerten eine einzelne positive Antwort zuweist.Für weitere Informationen siehe Erstellen eines kategorialen Responsefelds in Kapitel 6 auf S. 42.

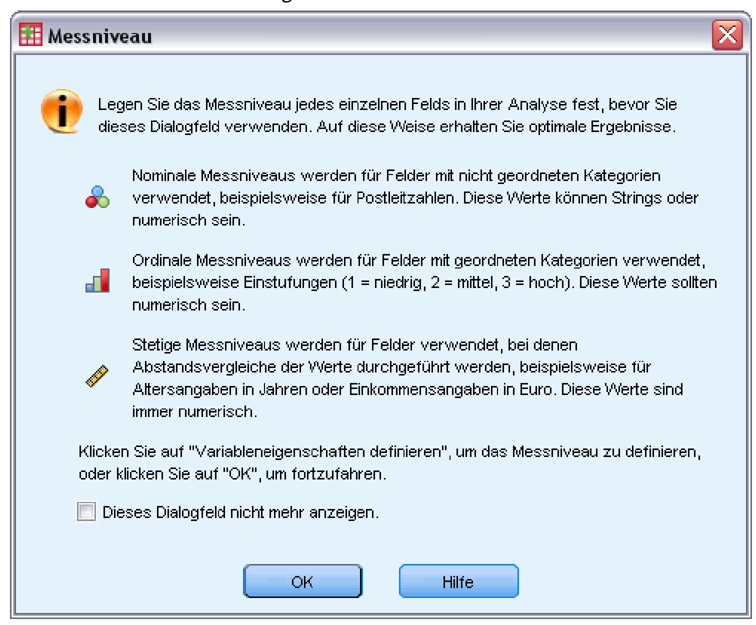
Durchführen der Analyse

▶ Um Neigungsbewertungen zu berechnen, wählen Sie in den Menüs folgende Optionen aus:

Option “Direct Marketing” (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Kontakte mit höchster Kaufneigung auswählen aus und klicken Sie auf Weiter.

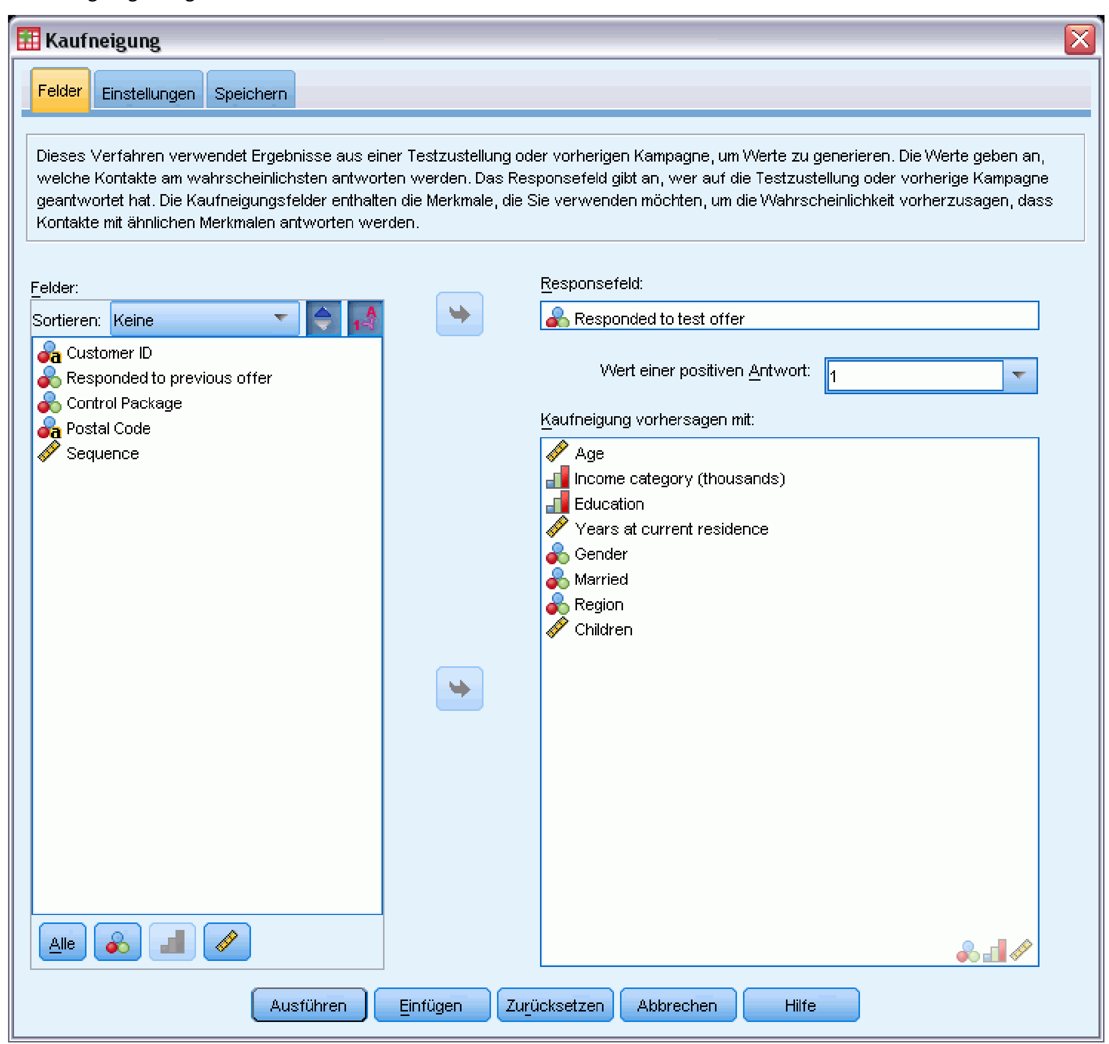
Abbildung 12-1
Messniveau-Warnmeldung



Die Messniveau-Warnmeldung wird angezeigt, weil das Messniveau die Berechnung der Ergebnisse für diese Prozedur beeinträchtigt. PASW Statistics weist neuen Daten-Sets anhand einiger einfacher Regeln automatisch ein Messniveau zu, wobei es sich hierbei nicht notwendigerweise um das beste Messniveau für alle Felder in Ihrem Daten-Set handeln muss. In diesem Beispiel-Daten-Set wurde allen Feldern das korrekte Messniveau zugewiesen.

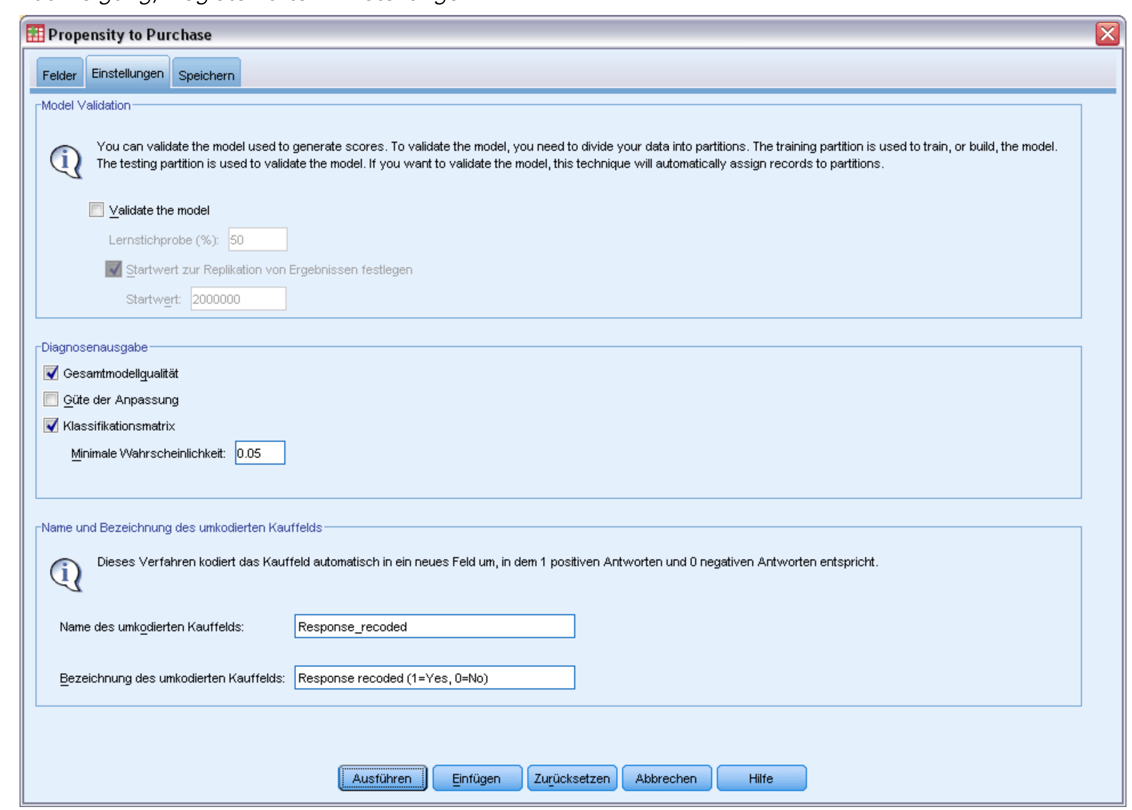
▶ Klicken Sie zum Fortfahren auf OK.

Abbildung 12-2
Kaufneigung, Registerkarte "Felder"



- Wählen Sie bei "Responsefeld" *Auf Testangebot geantwortet* aus.
- Wählen Sie bei "Wert für positive Antworten" *Ja* aus der Dropdown-Liste aus. Im Textfeld wird der Wert 1 angezeigt, da es sich bei "Ja" eigentlich um ein Wertelabel handelt, das zum aufgezeichneten Wert 1 gehört. (Wenn für den Wert für positive Antworten kein Wertelabel definiert wurde, können Sie den Wert einfach in das Textfeld eingeben.)
- Wählen Sie bei "Neigung vorhersagen durch" *Alter*, *Einkommensklasse*, *Schulbildung*, *Jahre an aktuellem Wohnort*, *Geschlecht*, *Verheiratet*, *Region* und *Kinder* aus.
- Klicken Sie auf die Registerkarte Einstellungen.

Abbildung 12-3
Kaufneigung, Registerkarte "Einstellungen"

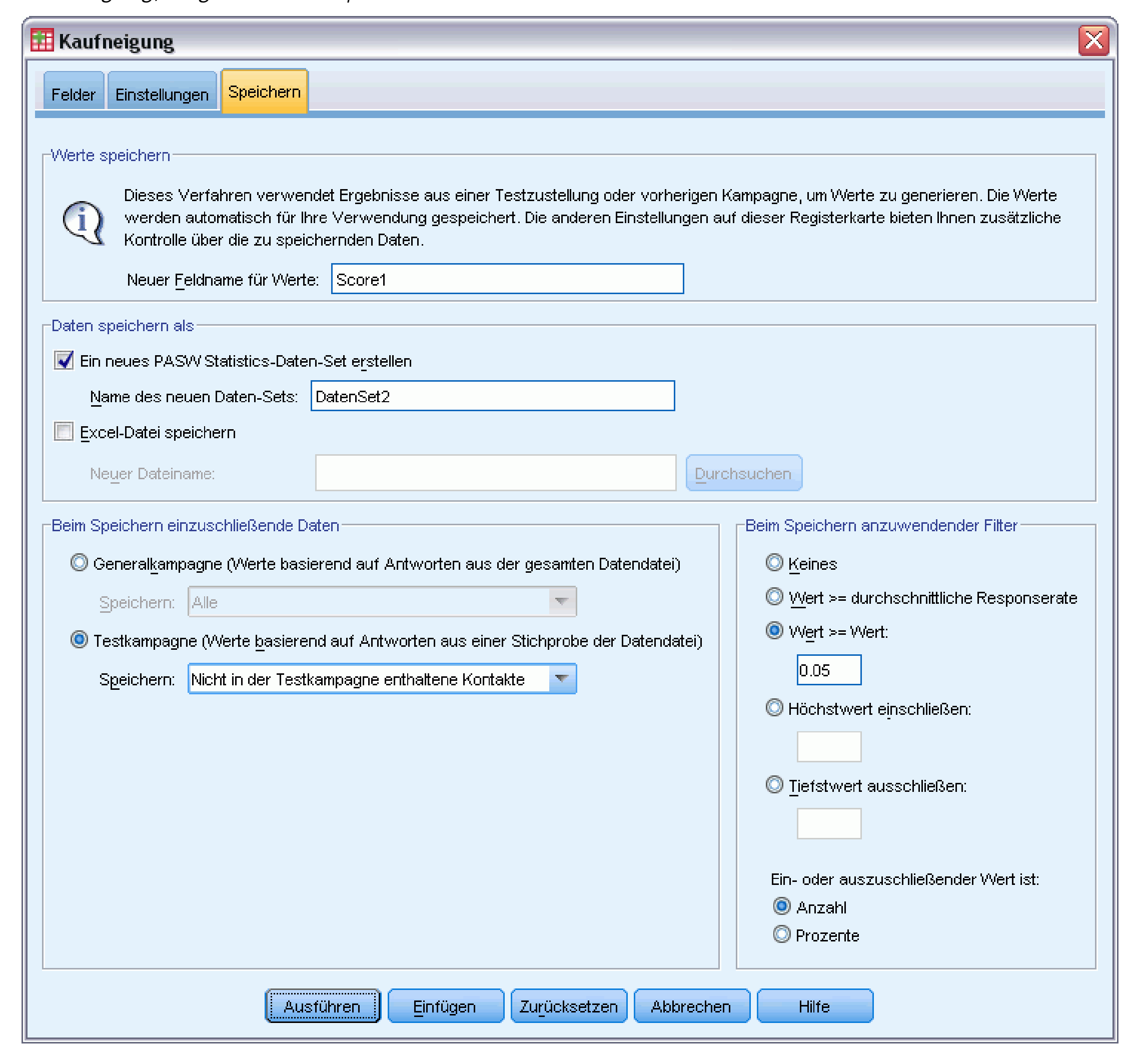


- Aktivieren Sie "Gesamtqualität" in der Gruppe "Diagnosenausgabe" und wählen Sie dann "Klassifikationsmatrix" aus.

- Geben Sie bei "Minimale Wahrscheinlichkeit" den Wert 0,05 ein. Allgemein sollten Sie einen Wert angeben, der in der Nähe Ihrer minimalen, als Anteil ausgedrückten Zielresponserate liegt. Ein Wert von 0,05 entspricht einer Responserate von 5 %.

- Klicken Sie anschließend auf die Registerkarte Speichern.

Abbildung 12-4
Kaufneigung, Registerkarte “Speichern”



- Aktivieren Sie in der Gruppe “Daten speichern als” “Neues PASW Statistics-Daten-Set erstellen”.
- Wählen Sie in der Gruppe “Beim Speichern einzuschließende Daten” “Testkampagne” (Bewertungen auf der Basis von Antworten aus einer Stichprobe der Datendatei) aus.
- Wählen Sie aus der Dropdown-Liste “Speichern” die Kontakte aus, die nicht in der Testkampagne enthalten waren.
- Wählen Sie in der Gruppe “Beim Speichern anzuwendender Filter” “Bewertung >= Wert” aus und geben Sie den Wert 0,05 ein.

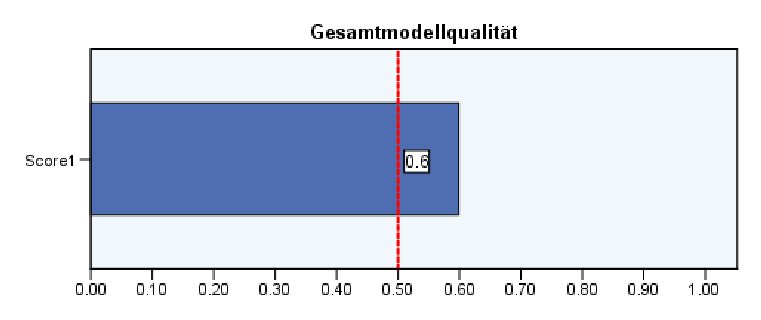
Die Bewertungen werden für alle Kontakte berechnet, einschließlich jener, die nicht in der Testkampagne enthalten waren. Das neue Daten-Set schließt Kontakte aus der Testkampagne aus (schließlich möchten Sie ihnen sicherlich das gleiche Angebot nicht noch einmal schicken) und schließt nur jene Kontakte ein, die nicht in der Testkampagne enthalten waren und eine Neigungsbewertung größer als 0,05 aufweisen. Eine Neigungsbewertung von 0,05 entspricht

einer vorhergesagten Wahrscheinlichkeit von 5 %. Wenn das Modell zuverlässig ist, können Sie demnach eine Responserate von 5 % aus der Liste der Kontakte im neuen Daten-Set erwarten.

- ▶ Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Ausgabe

Abbildung 12-5
Gesamtmodellqualität



Das Diagramm zur Gesamtmodellqualität bietet einen kurzen visuellen Überblick über die Qualität des Modells. In diesem Zusammenhang wird durch das Modell bestimmt, wie die Neigungsbewertungen berechnet werden. Wenn also das Modell von schlechter Qualität ist, sind die Bewertungen wahrscheinlich nicht sehr hilfreich. Bei einem guten Modell sollte der Wert der Gesamtmodellqualität über 0,5 liegen. In diesem Beispiel liegt der Wert bei 0,6, d. h. das Modell ist besser als die zufällige Vorhersage.

Abbildung 12-6
Klassifikationsmatrix

Klassifizierungstabelle[a]

			Vorhergesagt		
			Response recoded (1=Yes, 0=No)		Prozentsatz der Richtigen
	Beobachtet		No	Yes	
Schritt 1	Response recoded (1=Yes, 0=No)	No	1026	794	56.4
		Yes	27	70	72.2
	Gesamtprozentsatz				57.2

a.Der Trennwert lautet .050

Die Klassifikationsmatrix gibt ebenfalls Aufschluss über die Funktionsqualität des Modells, indem sie vorhergesagte Responseraten mit den tatsächlichen (beobachteten) Responseraten vergleicht. Die Gesamtrate für korrekte Vorhersagen liegt nur bei 57 %, die Rate für korrekte Vorhersagen bei positiven Antworten dagegen bei 72,2 %.

Diese Prozentwerte werden von der minimalen Wahrscheinlichkeit 0,05 beeinflusst, die auf der Registerkarte “Einstellungen” eingegeben wurde. Allen Kontakten in der Testkampagne mit einer Neigungsbewertung von 0,05 oder höher wird die vorhergesagte Antwortkategorie “Ja” zugewiesen. Die Neigungsbewertung ist die Wahrscheinlichkeit, dass der Kontakt positiv antworten wird, was bedeutet, dass jedem Kontakt, der mit einer Wahrscheinlichkeit von 5 % oder höher positiv antwortet, die vorhergesagte Kategorie “Ja” zugewiesen wird. Wenn das Modell von guter Qualität ist, sollten Sie daher eine Responserate von 5 % oder höher erzielen, wenn Ihre Zielgruppe Kontakte mit einer Neigungsbewertung von 0,05 oder höher ist.

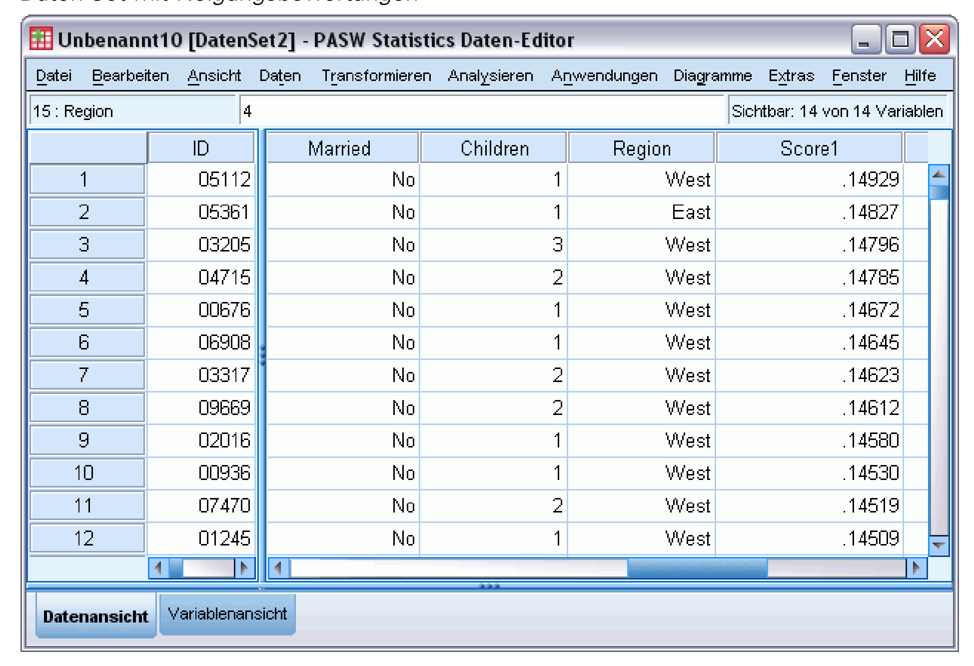
Wie gut war die korrekte Vorhersage positiver Antworten durch das Modell? Um dies herauszufinden, ist eine kleine zusätzliche Berechnung nötig. Die Gesamtanzahl der vorhergesagten “Ja”-Antworten ist 868 (798+70), wovon 70 korrekt klassifiziert wurden, sodass die korrekte Vorhersagerate für positive Antworten 70/868 oder 8 % beträgt.

Sie können eine deutlich höhere korrekte Gesamtklassifizierungsrate erzielen, indem Sie eine minimale Wahrscheinlichkeit von 0,5 verwenden, denn dadurch wird angenommen, dass die Wahrscheinlichkeit zwischen positiven und negativen Antworten gleich hoch ist. So würde sich allerdings eine Klassifikationsmatrix ergeben, in der allen Kontakten die vorhergesagte Kategorie “Nein” zugewiesen würde. Wenn die tatsächliche positive Responserate ca. 5 % beträgt, würde dies eine korrekte Gesamtklassifikationsrate von 95 % ergeben, jedoch eine korrekte Klassifikationsrate für positive Antworten von 0 %, was nicht sonderlich hilfreich wäre.

Anmerkung: Die von Ihnen für die Klassifikationsmatrix verwendete minimale Wahrscheinlichkeit hat keine Auswirkungen auf Bewertungen, die von der Prozedur erzeugt wurden, sondern nur auf die Diagnosenausgabe in der Klassifikationsmatrix.

Neigungsbewertungen

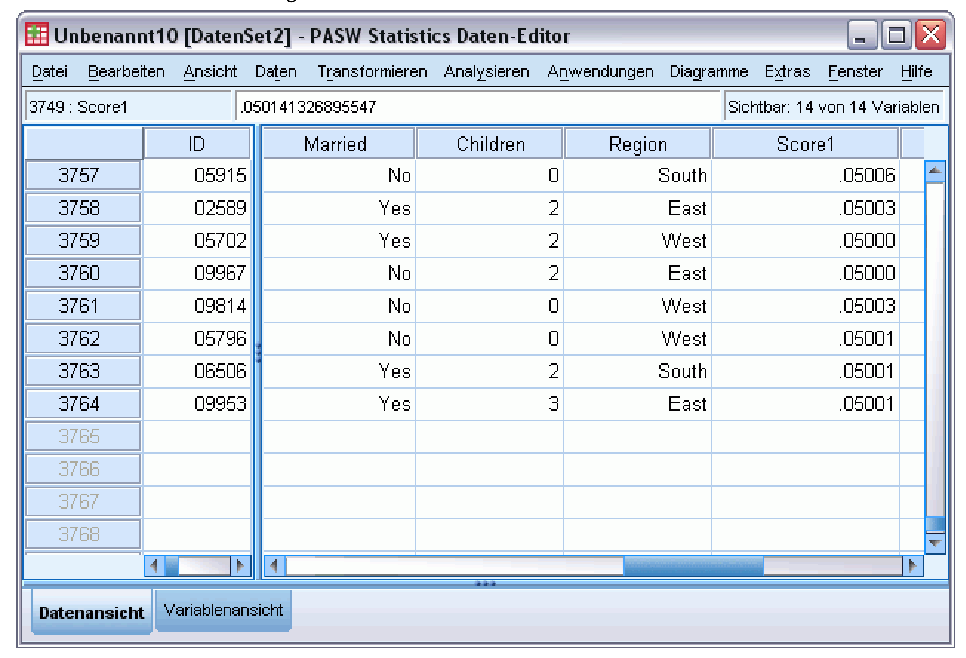
Abbildung 12-7
Daten-Set mit Neigungsbewertungen



Der Hauptzweck dieser Prozedur ist selbstverständlich das Erzeugen von Neigungsbewertungen. Für jeden Kontakt entspricht die Neigungsbewertung der als Anteil ausgedrückten Wahrscheinlichkeit, dass der Kontakt antworten wird. Zum Beispiel bedeutet eine Neigungsbewertung von 0,154, dass es eine Wahrscheinlichkeit von 15,4 Prozent gibt, dass der Kontakt antwortet. Diese Bewertungen werden automatisch dem Original-Daten-Set hinzugefügt, Sie können aber auch ein neues Daten-Set oder eine Excel-Datei erstellen, die nur jene Kontakte enthält, die bestimmte Kriterien erfüllen.

Neigungsbewertungen werden für alle Datensätze mit gültigen (nicht fehlenden) Werten für die ausgewählten Einflussvariablenfelder berechnet, selbst wenn das Modell für die Zuweisung der Bewertungen nur auf einer kleinen Stichprobe der Kontakte beruht. In diesem Beispiel basiert das Modell auf einer Testkampagne, die nur 20 % aller Kontakte im Daten-Set enthält. Anhand dieser Ergebnisse wurden die Bewertungen anschließend den restlichen Kontakten im Daten-Set zugeordnet.

Abbildung 12-8
Neues Daten-Set mit ausgewählten Kontakten



	ID	Married	Children	Region	Score1
3757	05915	No	0	South	.05006
3758	02589	Yes	2	East	.05003
3759	05702	Yes	2	West	.05000
3760	09967	No	2	East	.05000
3761	09814	No	0	West	.05003
3762	05796	No	0	West	.05001
3763	06506	Yes	2	South	.05001
3764	09953	Yes	3	East	.05001
3765					
3766					
3767					
3768					

In diesem Beispiel wurde ein neues Daten-Set erstellt, das in die Testkampagne aufgenommene Kontakte ausschließt (Sie möchten ihnen sicherlich das gleiche Angebot nicht noch einmal schicken) und auch Kontakte mit einer Neigungsbewertung unter 0,05 ausschließt, sodass eine Liste mit Kontakten entsteht, die nicht in die Testkampagne aufgenommen wurden und das Ergebnis der Gesamt-Responserate 5 % oder mehr betragen sollte.

Die Testkampagne enthielt ca. 20 % der insgesamt 10.000 Kontakte, sodass für die Folgekampagne rund 8.000 potenzielle Kontakte übrig blieben. Das Modell grenzt die Liste jedoch auf knapp unter 3.800 ein, wenn Sie eine Ziel-Responserate von mindestens 5 % erzielen möchten.

Anwendung des Bewertungsmodells auf ein anderes Daten-Set

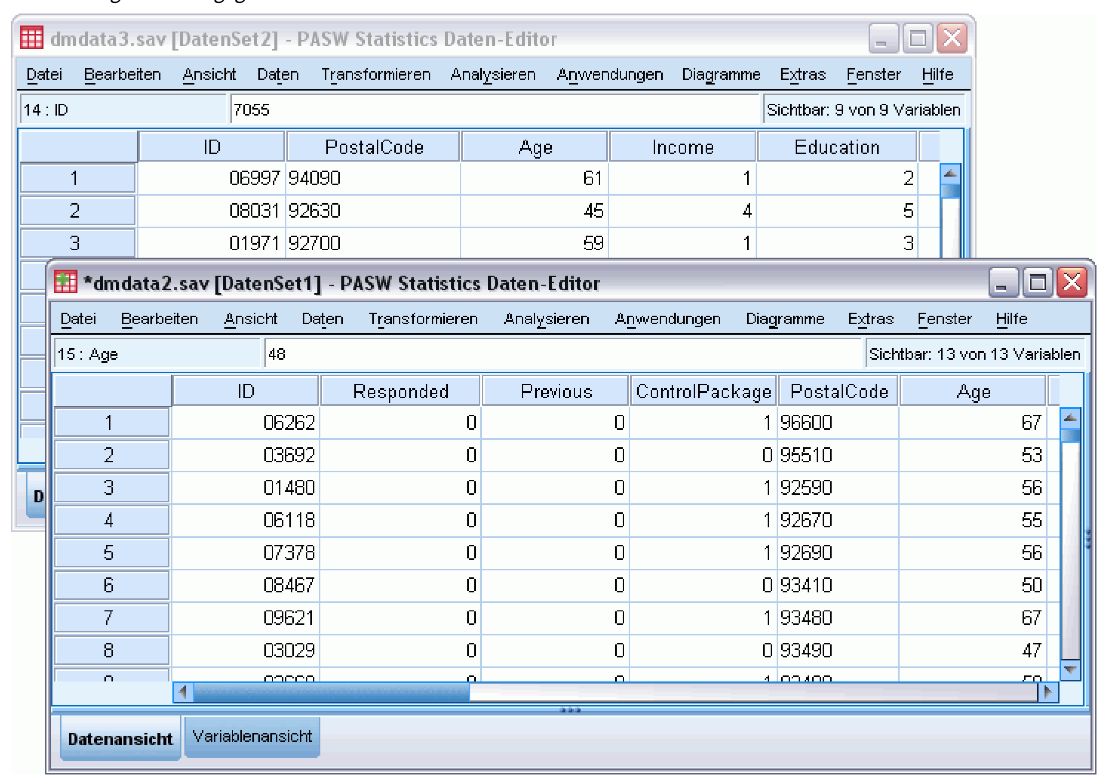
Neigungsbewertungen werden für alle Datensätze mit gültigen (nicht fehlenden) Werten für die ausgewählten Einflussvariablenfelder berechnet, selbst wenn das Modell für die Zuweisung der Bewertungen nur auf einer kleinen Stichprobe der Kontakte beruht. Kontakte, die nicht in der Testsendung enthalten waren, haben fehlende Werte im Responsefeld, sie erhalten aber dennoch Neigungsbewertungen, solange sie über nichtfehlende Werte für die ausgewählten Einflussvariablenfelder verfügen. Als einzige Einschränkung gilt, dass zum Zeitpunkt der Ausführung der Prozedur zur Berechnung der Bewertungen für alle Kontakte Datensätze im Daten-Set vorhanden sein müssen.

Doch was geschieht, wenn sich die zwei Kontaktgruppen nicht im selben Daten-Set befinden? Was passiert, wenn sich die Kontakte und Ergebnisse für den Test oder die frühere Kampagne in einer Datei und die Kontakte, die Sie bewerten möchten, in einer anderen Datei befinden?

Die Lösung für dieses Problem ist relativ einfach: Fügen Sie die beiden Dateien zusammen, bevor Sie die Prozedur ausführen.

Zusammenfügen von zwei Daten-Sets

Abbildung 12-9
Mehrere gleichzeitig geöffnete Daten-Sets



In diesem Beispiel werden zwei PASW Statistics-Datendateien zusammengefügt, es gelten jedoch die gleichen allgemeinen Regeln für das Zusammenfügen von Daten-Sets aus anderen Quellen wie Excel-Dateien oder Datenbanktabellen.

▶ Öffnen Sie die Datendateien *dmdata2.sav* und *dmdata3.sav*. Diese Dateien befinden sich im Ordner *Samples* im Installationsverzeichnis. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

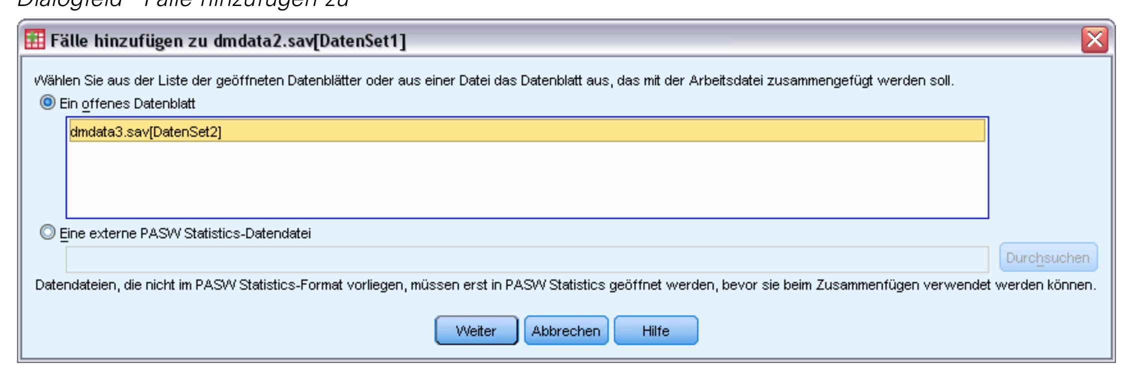
Die Datei *dmdata2.sav* enthält Informationen über die in der Testsendung enthaltenen Kontakte einschließlich der Ergebnisse aus der Testsendung. Die Datei *dmdata3.sav* enthält demografische Informationen über eine andere Kontaktgruppe.

▶ Stellen Sie sicher, dass das Daten-Editor-Fenster mit der Datei *dmdata2.sav* aktiv ist. (Klicken Sie auf einen beliebigen Punkt im Fenster, um es zu aktivieren.)

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

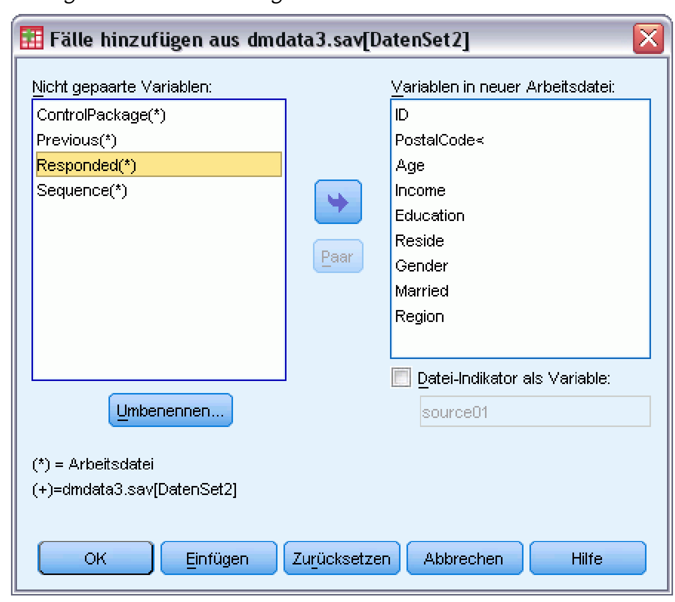
Daten
  Dateien zusammenfügen
    Fälle hinzufügen

Abbildung 12-10
Dialogfeld "Fälle hinzufügen zu"



▶ Wählen Sie ein geöffnetes Daten-Set und anschließend die Datei *dmdata3.sav* aus der Liste aus.

Abbildung 12-11
Dialogfeld "Fälle hinzufügen aus"



▶ Wählen Sie *Geantwortet* in der Liste "Nicht gepaarte Variablen" aus und verschieben Sie sie in die Liste "Variablen in neuem aktivem Daten-Set".

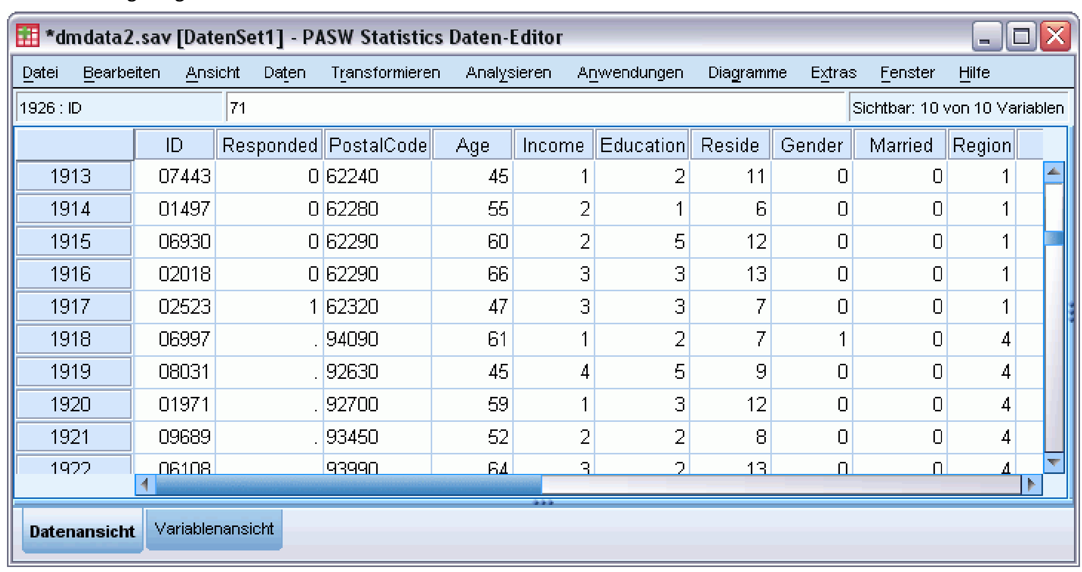
Die Liste "Nicht gepaarte Variablen" enthält standardmäßig Variablen (Felder), die nur in einem der beiden Daten-Sets vorhanden sind. Da nur ein Daten-Set Antwortdaten für die Testsendung enthält, befindet sich das Feld mit den Antwortergebnissen, *Geantwortet*, anfangs in der Liste "Nicht gepaarte Variablen"; das Feld wird jedoch benötigt, um die Prozedur auszuführen und Neigungsbewertungen zu berechnen.

Alle als Einflussvariablen verwendeten demografischen Felder sind in beiden Daten-Sets vorhanden, d. h. sie werden automatisch in die Liste der Variablen für das neue Daten-Set aufgenommen. Alle Felder, die Sie als Einflussvariablen verwenden möchten, müssen in beiden Daten-Sets vorhanden sein. Jedes Feld, das Sie aus der Liste "Nicht gepaarte Variablen"

hinzufügen, enthält fehlende Werte für alle Datensätze aus dem anderen Daten-Set, sodass sich für diese Datensätze fehlende Neigungsbewertungen ergeben, wenn Sie diese Felder als Einflussvariablen verwenden.

▶ Klicken Sie auf OK, um die beiden Daten-Sets zusammenzufügen.

Abbildung 12-12
Zusammengefügtes Daten-Set



	ID	Responded	PostalCode	Age	Income	Education	Reside	Gender	Married	Region
1913	07443	0	62240	45	1	2	11	0	0	1
1914	01497	0	62280	55	2	1	6	0	0	1
1915	06930	0	62290	60	2	5	12	0	0	1
1916	02018	0	62290	66	3	3	13	0	0	1
1917	02523	1	62320	47	3	3	7	0	0	1
1918	06997	.	94090	61	1	2	7	1	0	4
1919	08031	.	92630	45	4	5	9	0	0	4
1920	01971	.	92700	59	1	3	12	0	0	4
1921	09689	.	93450	52	2	2	8	0	0	4
1922	06108	.	93990	64	3	2	13	0	0	4

In dem zusammengefügten Daten-Set besitzen alle Datensätze aus der Datendatei, die nicht in der Testsendung enthalten waren, fehlende Werte für das Responsefeld. Dies stellt jedoch kein Problem dar, da für diese Datensätze trotzdem Bewertungen berechnet werden, solange diese nicht über fehlende Werte für eine der Einflussvariablen verfügen.

Nun können Sie die Kaufneigung zur Berechnung von Neigungsbewertungen verwenden. Für weitere Informationen siehe Durchführen der Analyse auf S. 86.

Zusammenfassung

Für die Kaufneigung werden Ergebnisse einer Testsendung oder einer früheren Kampagne verwendet, um Neigungsbewertungen zu erstellen. Die Bewertungen zeigen anhand von zahlreichen ausgewählten Merkmalen an, bei welchen Kontakten die Wahrscheinlichkeit einer Antwort am höchsten ist. Neigungsbewertungen werden für alle Datensätze mit gültigen (nicht fehlenden) Werten für die ausgewählten Einflussvariablenfelder berechnet, selbst wenn das Modell für die Zuweisung der Bewertungen nur auf einer kleinen Stichprobe der Kontakte beruht. Kontakte, die nicht in der Testsendung enthalten waren, haben fehlende Werte im Responsefeld, sie erhalten aber dennoch Neigungsbewertungen, solange sie über nichtfehlende Werte für die ausgewählten Einflussvariablenfelder verfügen. Wenn sich die Kontakte und Ergebnisse für den Test oder die frühere Kampagne in einer Datei und die Kontakte, die Sie bewerten möchten, in

einer anderen Datei befinden, fügen Sie die beiden Dateien zusammen, bevor Sie die Prozedur ausführen.

Kapitel
13

Kontrollpakettest

Dieses Verfahren vergleicht Marketingkampagnen, um herauszufinden, ob es hinsichtlich der Wirksamkeit signifikante Unterschiede zwischen verschiedenen Paketen oder Angeboten gibt. Die Kampagnenwirksamkeit wird anhand von Antworten gemessen.

Zum Beispiel möchte die Direktmarketing-Abteilung eines Unternehmens herausfinden, ob eine neue Verpackungsgestaltung mehr positive Antworten erzeugt als die bestehende Verpackung. Daher verschicken sie Testsendungen, um zu ermitteln, ob die neue Verpackung eine deutlich höhere positive Responserate erzeugt. Die Testsendung besteht aus einer Kontrollgruppe, die die aktuelle Verpackung erhält, und einer Testgruppe, an die die neue Verpackungsgestaltung geschickt wird. Die Ergebnisse der zwei Gruppen werden dann miteinander verglichen, um zu sehen, ob ein deutlicher Unterschied besteht.

Diese Informationen finden Sie in der Datei *dmdata.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 103.

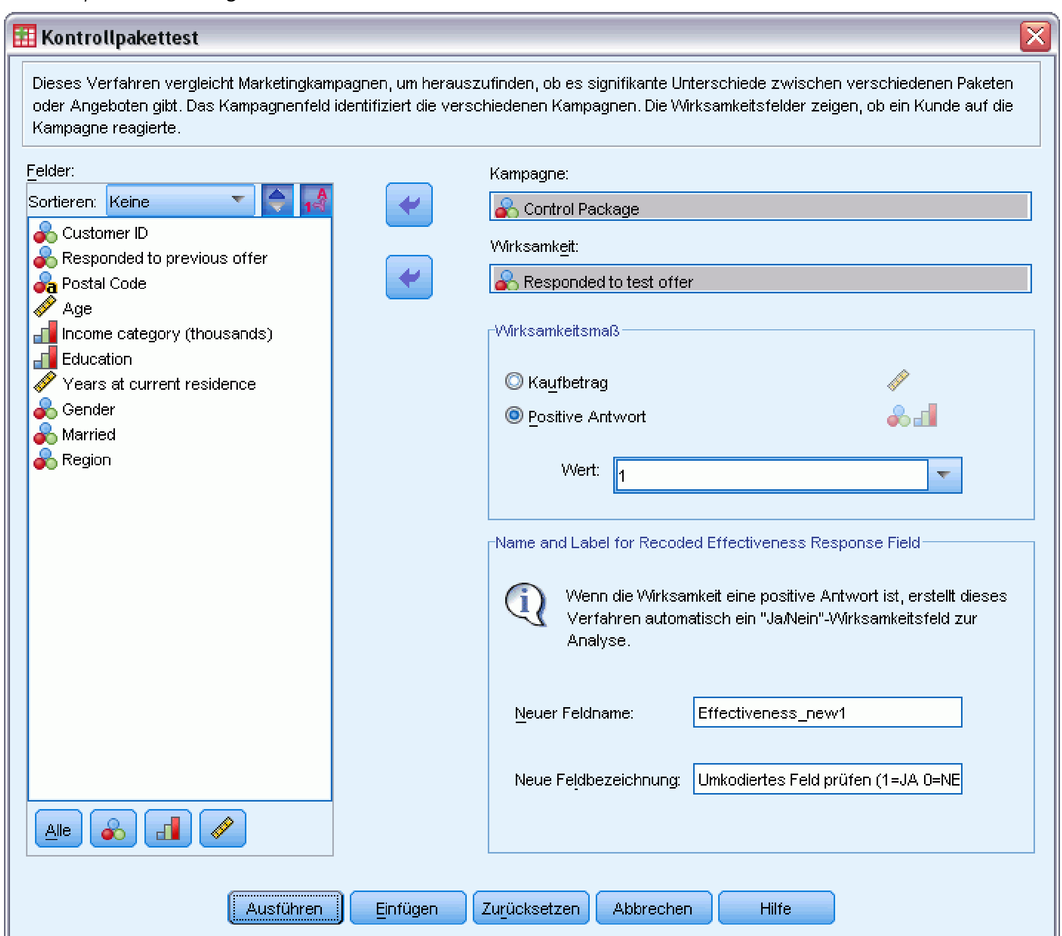
Durchführen der Analyse

▶ Zum Erstellen eines Kontrollpakettests wählen Sie die folgenden Befehle aus den Menüs aus:

Option "Direct Marketing" (Direktmarketing)
  Verfahren wählen

▶ Wählen Sie Wirksamkeit der Kampagnen vergleichen (Kontrollpakettest) und klicken Sie auf Weiter.

Abbildung 13-1
Kontrollpakettest, Registerkarte "Felder"



- Wählen Sie bei "Kampagnenfeld" *Kontrollpaket* aus.
- Wählen Sie bei "Wirksamkeits-Responsefeld" *Auf Testangebot geantwortet* aus.
- Wählen Sie "Antwort" aus.
- Wählen Sie bei "Wert für positive Antworten" *Ja* aus der Dropdown-Liste aus. Im Textfeld wird der Wert 1 angezeigt, da es sich bei "Ja" eigentlich um ein Wertelabel handelt, das zum aufgezeichneten Wert 1 gehört. (Wenn für den Wert für positive Antworten kein Wertelabel definiert wurde, können Sie den Wert einfach in das Textfeld eingeben.)

  Es wird automatisch ein neues Feld erstellt, in dem 1 positiven Antworten und 0 negativen Antworten entspricht; die Analyse wird in dem neuen Feld durchgeführt. Sie können den Standardnamen und die Standardbeschriftung durch eigene Angaben ersetzen. In diesem Beispiel wird der bereits angegebene Feldname verwendet.
- Klicken Sie auf Ausführen, um die Prozedur auszuführen.

Ausgabe

Abbildung 13-2
Ausgabe des Kontrollpakettests

		Control Package			
		Control		Test	
		Anzahl	Anzahl der Spalten (%)	Anzahl	Anzahl der Spalten (%)
Effectivenss(1=YES 0=NO)	0	875	96.2%	945	93.8%
	1	35	3.8%	62	6.2%

Vergleiche der Spaltenanteile[a]

		Control Package	
		Control	Test
		(A)	(B)
Effectivenss(1=YES 0=NO)	0	B	
	1		A

Die Ergebnisse beruhen auf zweiseitigen Tests mit einem Signifikanzniveau von 0.05. Für jedes signifikante Paar wird die Erläuterung der Kategorie mit den kleineren Spaltenanteilen unter der Kategorie mit den größeren Spaltenanteilen angezeigt.

a. Mit Hilfe der Bonferroni-Korrektur werden die Tests an alle paarweisen Vergleiche innerhalb einer Zeile der jeweils innersten Untertabelle angepaßt.

Die Ausgabe aus der Prozedur enthält eine Tabelle, in der Häufigkeiten und Prozentwerte von positiven und negativen Antworten für jede anhand des Kampagnenfelds definierte Gruppe sowie eine Tabelle, die aufzeigt, ob die Gruppen-Responseraten stark voneinander abweichen.

- Bei *Wirksamkeit* handelt es sich um die umkodierte Version des Responsefelds, in dem 1 positiven Antworten und 0 negativen Antworten entspricht.
- Die positive Responserate für das Kontrollpaket ist 3,8 % und die positive Responserate für das Testpaket ist 6,2 %.
- In der Tabelle für den Spaltenanteiletest sehen Sie, ob die Unterschiede zwischen den Gruppen stark voneinander abweichen. Jeder Kampagnengruppe wird ein anderer Buchstabe zugewiesen, in diesem Beispiel *A* für das Kontrollpaket und *B* für das Testpaket.
- Falls ein signifikanter Unterschied besteht, erscheint der Buchstabe für die Kategorie mit der kleineren Responserate in der Zelle für die Kategorie mit der größeren Responserate. Das heißt, das *A* in der Zelle, die den positiven Antworten für das Testpaket entspricht, zeigt, dass die positive Responserate für das Testpaket deutlich höher ist als die positive Responserate für die Kontrollgruppe. (Umgekehrt ist die negative Responserate für die Kontrollgruppe deutlich höher als die negative Responserate für die Testgruppe.)

Zusammenfassung

Der Kontrollpakettest vergleicht Marketingkampagnen, um herauszufinden, ob es hinsichtlich der Effektivität signifikante Unterschiede zwischen verschiedenen Paketen oder Angeboten gibt. In diesem Beispiel war der Wert der positiven Antworten für das Testpaket mit 6,2 % deutlich höher als die positive Responserate von 3,8 % für das Kontrollpaket. Dies lässt darauf schließen, dass Sie die neue Verpackungsgestaltung anstelle der alten verwenden sollten, aber Sie müssen

unter Umständen noch andere Faktoren berücksichtigen, so etwa zusätzliche Kosten, die das neue Verpackungsdesign verursacht.

Anhang



Beispieldateien

Die zusammen mit dem Produkt installierten Beispieldateien finden Sie im Unterverzeichnis *Samples* des Installationsverzeichnisses. Für jeder der folgenden Sprachen gibt es einen eigenen Ordner innerhalb des Unterverzeichnisses “Samples”: Englisch, Französisch, Deutsch, Italienisch, Japanisch, Koreanisch, Polnisch, Russisch, Vereinfachtes Chinesisch, Spanisch und Traditionelles Chinesisch.

Nicht alle Beispieldateien stehen in allen Sprachen zur Verfügung. Wenn eine Beispieldatei nicht in einer Sprache zur Verfügung steht, enthält der jeweilige Sprachordner eine englische Version der Beispieldatei.

Beschreibungen

Im Folgenden finden Sie Kurzbeschreibungen der in den verschiedenen Beispielen in der Dokumentation verwendeten Beispieldateien.

- **accidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die alters- und geschlechtsabhängige Risikofaktoren für Autounfälle in einer bestimmten Region untersucht. Jeder Fall entspricht einer Kreuzklassifikation von Alterskategorie und Geschlecht.
- **adl.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die Vorteile einer vorgeschlagenen Therapieform für Schlaganfallpatienten zu ermitteln. Ärzte teilten weibliche Schlaganfallpatienten nach dem Zufallsprinzip jeweils einer von zwei Gruppen zu. Die erste Gruppe erhielt die physische Standardtherapie, die zweite erhielt eine zusätzliche Emotionaltherapie. Drei Monate nach den Behandlungen wurden die Fähigkeiten der einzelnen Patienten, übliche Alltagsaktivitäten auszuführen, als ordinale Variablen bewertet.
- **advert.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Einzelhändlers geht, die Beziehungen zwischen den in Werbung investierten Beträgen und den daraus resultierenden Umsätzen zu untersuchen. Zu diesem Zweck hat er die Umsätze vergangener Jahre und die zugehörigen Werbeausgaben zusammengestellt.
- **aflatoxin.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests von Maisernten auf Aflatoxin geht, ein Gift, dessen Konzentration stark zwischen und innerhalb von Ernteerträgen schwankt. Ein Kornverarbeitungsbetrieb hat aus 8 Ernteerträgen je 16 Proben erhalten und das Aflatoxinniveau in Teilen pro Milliarde (parts per billion, PPB) gemessen.
- **aflatoxin20.sav.** Diese Datendatei enthält die Aflatoxinmessungen aus jeder der 16 Stichproben aus den Erträgen 4 und 8 der Datendatei *aflatoxin.sav*.

- **anorectic.sav.** Bei der Ausarbeitung einer standardisierten Symptomatologie anorektischen/bulimischen Verhaltens führten Forscher) eine Studie mit 55 Jugendlichen mit bekannten Ess-Störungen durch. Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung erhielten die Patienten Scores für jedes von 16 Symptomen. Die Symptomwerte fehlen für Patient 71 zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben.
- **autoaccidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Versicherungsanalysten geht, ein Modell zur Anzahl der Autounfälle pro Fahrer unter Berücksichtigung von Alter und Geschlecht zu erstellen. Jeder Fall stellt einen Fahrer dar und erfasst das Geschlecht des Fahrers, sein Alter in Jahren und die Anzahl der Autounfälle in den letzten fünf Jahren.
- **band.sav.** Diese Datendatei enthält die hypothetischen wöchentlichen Verkaufszahlen von CDs für eine Musikgruppe. Daten für drei mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **bankloan.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, den Anteil der nicht zurückgezahlten Kredite zu reduzieren. Die Datei enthält Informationen zum Finanzstatus und demografischen Hintergrund von 850 früheren und potenziellen Kunden. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Bei den letzten 150 Fällen handelt es sich um potenzielle Kunden, deren Kreditrisiko die Bank als gering oder hoch einstufen möchte.
- **bankloan_binning.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen zum Finanzstatus und demografischen Hintergrund von 5.000 früheren Kunden enthält.
- **behavior.sav.** In einem klassischen Beispiel () wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = “ausgesprochen angemessen” bis 9 = “ausgesprochen unangemessen” zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.
- **behavior_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine zweidimensionale Lösung für *behavior.sav*.
- **brakes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik geht, die Scheibenbremsen für Hochleistungsautomobile herstellt. Die Datendatei enthält Messungen des Durchmessers von 16 Scheiben aus 8 Produktionsmaschinen. Der Zieldurchmesser für die Scheiben ist 322 Millimeter.
- **breakfast.sav.** In einer klassischen Studie () wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = “am meisten bevorzugt” bis 15 = “am wenigsten bevorzugt” zu ordnen. Die Bevorzugungen wurden in sechs unterschiedlichen Szenarien erfasst, von “Overall preference” (Allgemein bevorzugt) bis “Snack, with beverage only” (Imbiss, nur mit Getränk).
- **breakfast-overall.sav.** Diese Datei enthält die Daten zu den bevorzugten Frühstücksartikeln, allerdings nur für das erste Szenario, “Overall preference” (Allgemein bevorzugt).
- **broadband_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Anzahl der Abonnenten eines Breitband-Service, nach Region geordnet, enthält. Die Datendatei enthält die monatlichen Abonnentenzahlen für 85 Regionen über einen Zeitraum von vier Jahren.

- **broadband_2.sav.** Diese Datendatei stimmt mit *broadband_1.sav* überein, enthält jedoch Daten für weitere drei Monate.
- **car_insurance_claims.sav.** Ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set bezieht sich auf Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gamma-Verteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination aus Alter des Versicherungsnehmers, Fahrzeugtyp und Fahrzeugalter in Bezug zu setzen. Die Anzahl der eingereichten Schadensansprüche kann als Skalierungsgewicht verwendet werden.
- **car_sales.sav.** Diese Datendatei enthält hypothetische Verkaufsschätzer, Listenpreise und physische Spezifikationen für verschiedene Fahrzeugfabrikate und -modelle. Die Listenpreise und physischen Spezifikationen wurden von *edmunds.com* und Hersteller-Websites entnommen.
- **car_sales_uprepared.sav.** Hierbei handelt es sich um eine modifizierte Version der Datei *car_sales.sav*, die keinerlei transformierte Versionen der Felder enthält.
- **carpet.sav.** In einem beliebten Beispiel möchte einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Bevorzugung durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel, *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung setzt sich aus drei Faktorenebenen zusammen, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Ebenen (Nein oder Ja) für die letzten beiden Faktoren. 10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Bevorzugung. Diese Variable gibt ein Gesamtmaß der Bevorzugung für die Profile an.
- **carpet_prefs.sav.** Diese Datendatei beruht auf denselben Beispielen, wie für *carpet.sav* beschrieben, enthält jedoch die tatsächlichen Einstufungen durch jeden der 10 Kunden. Die Kunden wurden gebeten, die 22 Produktprofile in der Reihenfolge ihrer Präferenzen einzustufen. Die Variablen *PREF1* bis *PREF22* enthalten die IDs der zugeordneten Profile, wie in *carpet_plan.sav* definiert.
- **catalog.sav.** Diese Datendatei enthält hypothetische monatliche Verkaufszahlen für drei Produkte, die von einem Versandhaus verkauft werden. Daten für fünf mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **catalog_seasfac.sav.** Diese Datendatei ist mit *catalog.sav* identisch, außer, dass ein Set von saisonalen Faktoren, die mithilfe der Prozedur “Saisonale Zerlegung” berechnet wurden, sowie die zugehörigen Datumsvariablen hinzugefügt wurden.
- **cellular.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Mobiltelefonunternehmens geht, die Kundenabwanderung zu verringern. Scores für die Abwanderungsneigung (von 0 bis 100) werden auf die Kunden angewendet. Kunden mit einem Score von 50 oder höher streben vermutlich einen Anbieterwechsel an.
- **ceramics.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Herstellers geht, der ermitteln möchte, ob ein neue, hochwertige Keramiklegierung eine größere Hitzebeständigkeit aufweist als eine Standardlegierung. Jeder Fall entspricht einem Test einer der Legierungen; die Temperatur, bei der das Keramikwälzlager versagte, wurde erfasst.

- **cereal.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Umfrage geht, bei der 880 Personen nach ihren Frühstückgewohnheiten befragt wurden. Außerdem wurden Alter, Geschlecht, Familienstand und Vorliegen bzw. Nichtvorliegen eines aktiven Lebensstils (auf der Grundlage von mindestens zwei Trainingseinheiten pro Woche) erfasst. Jeder Fall entspricht einem Teilnehmer.
- **clothing_defects.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Bekleidungsfabrik geht. Aus jeder in der Fabrik produzierten Charge entnehmen die Kontrolleure eine Stichprobe an Bekleidungsartikeln und zählen die Anzahl der Bekleidungsartikel die inakzeptabel sind.
- **coffee.sav.** Diese Datendatei enthält Daten zum wahrgenommenen Image von sechs Eiskaffeemarken (). Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als “AA”, “BB”, “CC”, “DD”, “EE” und “FF” bezeichnet, um Vertraulichkeit zu gewährleisten.
- **contacts.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Kontaktlisten einer Gruppe von Vertretern geht, die Computer an Unternehmen verkaufen. Die einzelnen Kontaktpersonen werden anhand der Abteilung, in der sie in ihrem Unternehmen arbeiten und anhand ihrer Stellung in der Unternehmenshierarchie in Kategorien eingeteilt. Außerdem werden der Betrag des letzten Verkaufs, die Zeit seit dem letzten Verkauf und die Größe des Unternehmens, in dem die Kontaktperson arbeitet, aufgezeichnet.
- **creditpromo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Kaufhauses geht, die Wirksamkeit einer kürzlich durchgeführten Kreditkarten-Werbeaktion einzuschätzen. Dazu wurden 500 Karteninhaber nach dem Zufallsprinzip ausgewählt. Die Hälfte erhielt eine Werbebeilage, die einen reduzierten Zinssatz für Einkäufe in den nächsten drei Monaten ankündigte. Die andere Hälfte erhielt eine Standard-Werbebeilage.
- **customer_dbase.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, das die Informationen in seinem Data Warehouse nutzen möchte, um spezielle Angebote für Kunden zu erstellen, die mit der größten Wahrscheinlichkeit darauf ansprechen. Nach dem Zufallsprinzip wurde eine Untergruppe des Kundenstamms ausgewählt. Diese Gruppe erhielt die speziellen Angebote und die Reaktionen wurden aufgezeichnet.
- **customer_information.sav.** Eine hypothetische Datendatei mit Kundenmailingdaten wie Name und Adresse.
- **customers_model.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, auf die sich eine Marketingkampagne richtete. Zu diesen Daten gehören demografische Informationen, eine Übersicht über die bisherigen Einkäufe und die Angabe ob die einzelnen Personen auf die Kampagne ansprachen oder nicht. Jeder Fall entspricht einer Einzelperson.
- **customers_new.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, die potenzielle Kandidaten für Marketingkampagnen sind. Zu diesen Daten gehören demografische Informationen und eine Übersicht über die bisherigen Einkäufe für jede Person. Jeder Fall entspricht einer Einzelperson.
- **debate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die gepaarte Antworten auf eine Umfrage unter den Zuhörern einer politischen Debatte enthält (Antworten vor und nach der Debatte). Jeder Fall entspricht einem Befragten.

- **debate_aggregate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der die Antworten aus *debate.sav* aggregiert wurden. Jeder Fall entspricht einer Kreuzklassifikation der bevorzugten Politiker vor und nach der Debatte.
- **demo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Kundendatenbank geht, die zum Zwecke der Zusendung monatlicher Angebote erworben wurde. Neben verschiedenen demografischen Informationen ist erfasst, ob der Kunde auf das Angebot geantwortet hat.
- **demo_cs_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den ersten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einer anderen Stadt. Außerdem sind IDs für Region, Provinz, Landkreis und Stadt erfasst.
- **demo_cs_2.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den zweiten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einem anderen Stadtteil aus den im ersten Schritt ausgewählten Städten. Außerdem sind IDs für Region, Provinz, Landkreis, Stadt, Stadtteil und Wohneinheit erfasst. Die Informationen zur Stichprobenziehung aus den ersten beiden Stufen des Stichprobenplans sind ebenfalls enthalten.
- **demo_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfrageinformationen enthält die mit einem komplexen Stichprobenplan erfasst wurden. Jeder Fall entspricht einer anderen Wohneinheit. Es sind verschiedene Informationen zum demografischen Hintergrund und zur Stichprobenziehung erfasst.
- **dmdata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen über Demografie und Einkäufe für ein Direktmarketing-Unternehmen enthält.
- **dietstudy.sav.** Diese hypothetische Datendatei enthält die Ergebnisse einer Studie der "Stillman-Diät" . Jeder Fall entspricht einem Teilnehmer und enthält dessen Gewicht vor und nach der Diät in amerikanischen Pfund sowie mehrere Messungen des Triglyceridspiegels (in mg/100 ml).
- **dischargedata.sav.** Hierbei handelt es sich um eine Datendatei zum Thema *Seasonal Patterns of Winnipeg Hospital Use*, (Saisonale Muster der Belegung im Krankenhaus von Winnipeg) vom Manitoba Centre for Health Policy.
- **dvdplayer.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Entwicklung eines neuen DVD-Spielers geht. Mithilfe eines Prototyps hat das Marketing-Team Zielgruppendaten erfasst. Jeder Fall entspricht einem befragten Benutzer und enthält demografische Daten zu dem Benutzer sowie dessen Antworten auf Fragen zum Prototyp.
- **flying.sav.** Diese Datendatei enthält die Flugmeilen zwischen zehn Städten in den USA.
- **german_credit.sav.** Diese Daten sind aus dem Daten-Set "German credit" im Repository of Machine Learning Databases () an der Universität von Kalifornien in Irvine entnommen.
- **grocery_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *grocery_coupons.sav*, wobei die wöchentlichen Einkäufe zusammengefasst sind, sodass jeder Fall einem anderen Kunden entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und der verzeichnete ausgegebene Betrag ist nun die Summe der Beträge, die in den vier Wochen der Studie ausgegeben wurden.

- **grocery_coupons.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfragedaten enthält, die von einer Lebensmittelkette erfasst wurden, die sich für die Kaufgewohnheiten ihrer Kunden interessiert. Jeder Kunde wird über vier Wochen beobachtet, und jeder Fall entspricht einer Kundenwoche und enthält Informationen zu den Geschäften, in denen der Kunde einkauft sowie zu anderen Merkmalen, beispielsweise welcher Betrag in der betreffenden Woche für Lebensmittel ausgegeben wurde.
- **guttman.sav.** Bell () legte eine Tabelle zur Darstellung möglicher sozialer Gruppen vor. Guttman () verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: "crowds" (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), "audience" (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), "public" (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), "mobs" (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), "primary groups" (Primärgruppen, vertraulich), "secondary groups" (Sekundärgruppen, freiwillig) und "modern community" (die moderne Gesellschaft, ein lockerer Zusammenschluss, der aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).
- **healthplans.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Versicherungsgruppe geht, vier verschiedene Pläne zur Gesundheitsvorsorge für Kleinbetriebe zu evaluieren. Zwölf Inhaber von Kleinbetrieben (Arbeitgeber) wurden gebeten, die Pläne danach in eine Rangfolge zu bringen, wie gern sie sie ihren Mitarbeitern anbieten würden. Jeder Fall entspricht einem Arbeitgeber und enthält die Reaktionen auf die einzelnen Pläne.
- **health_funding.sav.** Hierbei handelt es sich um eine hypothetische Datei, die Daten zur Finanzierung des Gesundheitswesens (Betrag pro 100 Personen), Krankheitsraten (Rate pro 10.000 Personen der Bevölkerung) und Besuche bei medizinischen Einrichtungen/Ärzten (Rate pro 10.000 Personen der Bevölkerung) enthält. Jeder Fall entspricht einer anderen Stadt.
- **hivassay.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Bemühungen eines pharmazeutischen Labors, einen Schnelltest zur Erkennung von HIV-Infektionen zu entwickeln. Die Ergebnisse des Tests sind acht kräftiger werdende Rotschattierungen, wobei kräftigeren Schattierungen auf eine höhere Infektionswahrscheinlichkeit hindeuten. Bei 2.000 Blutproben, von denen die Hälfte mit HIV infiziert war, wurde ein Labortest durchgeführt.
- **hourlywagedata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zum Stundenlohn von Pflegepersonal in Praxen und Krankenhäusern mit unterschiedlich langer Berufserfahrung.
- **insurance_claims.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen möchte. Jeder Fall entspricht einem Anspruch.
- **insure.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die die Risikofaktoren untersucht, die darauf hinweisen, ob ein Kunde die Leistungen einer mit einer Laufzeit von 10 Jahren abgeschlossenen Lebensversicherung in Anspruch nehmen wird. Jeder Fall in der Datendatei entspricht einem Paar von Verträgen, je einer mit Leistungsforderung und der andere ohne, wobei die beiden Versicherungsnehmer in Alter und Geschlecht übereinstimmen.

- **judges.sav.** Hierbei handelt es sich um eine hypothetische Datendatei mit den Wertungen von ausgebildeten Kampfrichtern (sowie eines Sportliebhabers) zu 300 Kunstturnleistungen. Jede Zeile stellt eine Leistung dar; die Kampfrichter bewerteten jeweils dieselben Leistungen.
- **kinship_dat.sav.** Rosenberg und Kim haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach einem anderen Kriterium erfolgen als die erste. So wurden insgesamt sechs “Quellen” erzielt. Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15×15 Elementen. Die Anzahl der Zellen ist dabei gleich der Anzahl der Personen in einer Quelle minus der Anzahl der gemeinsamen Platzierungen der Objekte in dieser Quelle.
- **kinship_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine dreidimensionale Lösung für *kinship_dat.sav*.
- **kinship_var.sav.** Diese Datendatei enthält die unabhängigen Variablen *gender* (Geschlecht), *gener*(Generation) und *degree* (Verwandtschaftsgrad), die zur Interpretation der Dimensionen einer Lösung für *kinship_dat.sav* verwendet werden können. Insbesondere können sie verwendet werden, um den Lösungsraum auf eine lineare Kombination dieser Variablen zu beschränken.
- **mailresponse.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines Bekleidungsherstellers geht, der ermitteln möchte, ob die Verwendung von Briefsendungen für das Direktmarketing zu schnelleren Antworten führt als Postwurfsendungen. Die Mitarbeiter in der Bestellannahme erfassen, wie vielen Wochen nach der Postsendung die einzelnen Bestellungen aufgegeben wurden.
- **marketvalues.sav.** Diese Datendatei betrifft Hausverkäufe in einem Neubaugebiet in Algonquin, Illinois, in den Jahren 1999–2000. Diese Verkäufe sind in Grundbucheinträgen dokumentiert.
- **mutualfund.sav.** Diese Datendatei betrifft Aktienmarktdaten für verschiedene Technologieaktien, die in im Index S&P 500 verzeichnet sind. Jeder Fall entspricht einem Unternehmen.
- **nhis2000_subset.sav.** Die “National Health Interview Survey (NHIS)” ist eine große, bevölkerungsbezogene Umfrage in unter der US-amerikanischen Zivilbevölkerung. Es werden persönliche Interviews in einer landesweit repräsentativen Stichprobe von Haushalten durchgeführt. Für die Mitglieder jedes Haushalts werden demografische Informationen und Beobachtungen zum Gesundheitsverhalten und Gesundheitsstatus eingeholt. Diese Datendatei enthält eine Teilmenge der Informationen aus der Umfrage des Jahres 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Datendatei und Dokumentation öffentlich zugänglich. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Zugriff erfolgte 2003.
- **ozone.sav.** Die Daten enthalten 330 Beobachtungen zu sechs meteorologischen Variablen zur Vorhersage der Ozonkonzentration aus den übrigen Variablen. Bei früheren Untersuchungen (,) fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.

- **pain_medication.sav.** Diese hypothetische Datendatei enthält die Ergebnisse eines klinischen Tests für ein entzündungshemmendes Medikament zur Schmerzbehandlung bei chronischer Arthritis. Von besonderem Interesse ist die Zeitdauer, bis die Wirkung des Medikaments einsetzt und wie es im Vergleich mit bestehenden Medikamenten abschneidet.
- **patient_los.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen zu Patienten, die wegen des Verdachts auf Herzinfarkt in das Krankenhaus eingeliefert wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **patlos_sample.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen für eine Stichprobe von Patienten, denen während der Behandlung eines Herzinfarkts Thrombolytika verabreicht wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **polishing.sav.** Hierbei handelt es sich um die Datendatei "Nambeware Polishing Times" aus der Data and Story Library. Sie bezieht sich auf die Bemühungen eines Herstellers von Metallgeschirr (Nambe Mills, Santa Fe, New Mexico) zur zeitlichen Planung seiner Produktion. Jeder Fall entspricht einem anderen Artikel in der Produktpalette. Für jeden Artikel sind Durchmesser, Polierzeit, Preis und Produkttyp erfasst.
- **poll_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die öffentliche Unterstützung für einen Gesetzentwurf zu ermitteln, bevor er im Parlament eingebracht wird. Die Fälle entsprechen registrierten Wählern. Für jeden Fall sind County, Gemeinde und Wohnviertel des Wählers erfasst.
- **poll_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *poll_cs.sav* aufgeführten Wähler. Die Stichprobe wurde gemäß dem in der Plandatei *poll.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Beachten Sie jedoch Folgendes: Da im Stichprobenplan die PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*poll_jointprob.sav*). Die zusätzlichen Variablen zum demografischen Hintergrund der Wähler und ihrer Meinung zum vorgeschlagenen Gesetzentwurf wurden nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **property_assess.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen Bezirk (County) zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien, die im vergangenen Jahr in dem betreffenden County verkauft wurden. Jeder Fall in der Datendatei enthält die Gemeinde, in der sich die Immobilie befindet, den Bewerter, der die Immobilie besichtigt hat, die seit dieser Bewertung verstrichene Zeit, den zu diesem Zeitpunkt ermittelten Wert sowie den Verkaufswert der Immobilie.
- **property_assess_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen US-Bundesstaat zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien in dem betreffenden Bundesstaat. Jeder Fall in der Datendatei enthält das County, die Gemeinde und das Wohnviertel, in dem sich die Immobilie befindet, die seit der letzten Bewertung verstrichene Zeit sowie zu diesem Zeitpunkt ermittelten Wert.

- **property_assess_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *property_assess_cs.sav* aufgeführten Immobilien. Die Stichprobe wurde gemäß dem in der Plandatei *property_assess.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Die zusätzliche Variable *Current value* (Aktueller Wert) wurde nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **recidivism.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Zeit bis zu seiner zweiten Festnahme, sofern diese innerhalb von zwei Jahren nach der ersten Festnahme erfolgte.
- **recidivism_cs_sample.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter, der im Juni 2003 erstmals aus der Haft entlassen wurde, und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Daten zu seiner zweiten Festnahme, sofern diese bis Ende Juni 2006 erfolgte. Die Straftäter wurden aus per Stichprobenziehung ermittelten Polizeidirektionen ausgewählt (gemäß dem in *recidivism_cs.csplan* angegebenen Stichprobenplan). Da hierbei eine PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Eine hypothetische Datendatei mit Kauftransaktionsdaten wie Kaufdatum, gekauften Artikeln und Geldbetrag für jede Transaktion.
- **salesperformance.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung von zwei neuen Verkaufsschulungen geht. 60 Mitarbeiter, die in drei Gruppen unterteilt sind, erhalten jeweils eine Standardschulung. Zusätzlich erhält Gruppe 2 eine technische Schulung und Gruppe 3 eine Praxisschulung. Die einzelnen Mitarbeiter wurden am Ende der Schulung einem Test unterzogen und die erzielten Punkte wurden erfasst. Jeder Fall in der Datendatei stellt einen Lehrgangsteilnehmer dar und enthält die Gruppe, der der Lehrgangsteilnehmer zugeteilt wurde sowie die von ihm in der Prüfung erreichte Punktzahl.
- **satisf.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Zufriedenheitsumfrage, die von einem Einzelhandelsunternehmen in 4 Filialen durchgeführt wurde. Insgesamt wurden 582 Kunden befragt. Jeder Fall gibt die Antworten eines einzelnen Kunden wieder.
- **screws.sav.** Diese Datendatei enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln ().
- **shampoo_ph.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik für Haarpflegeprodukte geht. In regelmäßigen Zeitabständen werden Messwerte von sechs separaten Ausgangschargen erhoben und ihr pH-Wert erfasst. Der Zielbereich ist 4,5–5,5.
- **ships.sav.** Ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set bezieht sich auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe von Schiffstyp, Konstruktionszeitraum und Betriebszeitraum gemäß einer Poisson-Rate modelliert werden. Das Aggregat der Betriebsmonate für jede Zelle

der durch die Kreuzklassifizierung der Faktoren gebildeten Tabelle gibt die Werte für die Risikoanfälligkeit an.

- **site.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, neue Standorte für die betriebliche Expansion auszuwählen. Das Unternehmen beauftragte zwei Berater unabhängig voneinander mit der Bewertung der Standorte. Neben einem umfassenden Bericht gaben die Berater auch eine zusammenfassende Wertung für jeden Standort als “good” (gut) “fair” (mittelmäßig) oder “poor” (schlecht) ab.
- **siteratings.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Betatests der neuen Website eines E-Commerce-Unternehmens geht. Jeder Fall entspricht einem Beta-Tester, der die Brauchbarkeit der Website auf einer Skala von 0 bis 20 bewertete.
- **smokers.sav.** Diese Datendatei wurde aus der Umfrage “National Household Survey of Drug Abuse” aus dem Jahr 1998 abstrahiert und stellt eine Wahrscheinlichkeitsstichprobe US-amerikanischer Haushalte dar. Daher sollte der erste Schritt bei der Analyse dieser Datendatei darin bestehen, die Daten entsprechend den Bevölkerungstrends zu gewichten.
- **smoking.sav.** Hierbei handelt es sich um eine von Greenacre () vorgestellte hypothetische Tabelle. Die relevante Tabelle wird durch eine Kreuztabelle der Rauchgewohnheiten und der Berufskategorie gebildet. Die Variable *Berufsgruppe* enthält die Berufskategorien *Senior Manager*, *Junior Manager*, *Angestellter mit Erfahrung*, *Angestellter ohne Erfahrung* und *Sekretariat* sowie die Kategorie *National Average*, die als Ergänzung der Analyse dienen kann. Die Variable *Rauchen* enthält die Rauchgewohnheiten *Nichtraucher*, *Leicht*, *Mittel* und *Stark* sowie die Kategorien *No Alcohol* und *Alcohol*, die als Ergänzung der Analyse dienen können.
- **storebrand.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Verkaufsleiterin in einem Lebensmittelmarkt geht, die die Verkaufszahlen des Waschmittels der Eigenmarke gegenüber den anderen Marken steigern möchte. Sie erarbeitet eine Werbeaktion im Geschäft und spricht an der Kasse mit Kunden. Jeder Fall entspricht einem Kunden.
- **stores.sav.** Diese Datendatei enthält hypothetische monatliche Marktanteilsdaten für zwei konkurrierende Lebensmittelgeschäfte. Jeder Fall entspricht den Marktanteilsdaten für einen bestimmten Monat.
- **stroke_clean.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozeduren in der Option “Data Preparation” bereinigt wurde.
- **stroke_invalid.sav.** Diese hypothetische Datendatei enthält den ursprünglichen Zustand einer medizinischen Datenbank, der mehrere Dateneingabefehler aufweist.
- **stroke_survival.** In dieser hypothetischen Datendatei geht es um die Überlebenszeiten von Patienten, die nach einem Rehabilitationsprogramm wegen eines ischämischen Schlaganfalls mit einer Reihe von Problemen zu kämpfen haben. Nach dem Schlaganfall werden das Auftreten von Herzinfarkt, ischämischem Schlaganfall und hämorrhagischem Schlaganfall sowie der Zeitpunkt des Ereignisses aufgezeichnet. Die Stichprobe ist auf der linken Seite abgeschnitten, da sie nur Patienten enthält, die bis zum Ende des Rehabilitationprogramms, das nach dem Schlaganfall durchgeführt wurde, überlebten.
- **stroke_valid.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozedur “Daten validieren” überprüft wurde. Sie enthält immer noch potenziell anomale Fälle.

- **survey_sample.sav.** Diese hypothetische Datendatei enthält Umfragedaten einschließlich demografischer Daten und verschiedener Meinungskennzahlen.
- **tastetest.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung der Auswirkungen der Mulchfarbe auf den Geschmack von Pflanzenprodukten geht. Der Geschmack von Erdbeeren, die in rotem, blauem und schwarzem Rindenmulch gezogen wurden, wurde von Testpersonen auf einer ordinalen Skala (weit unter bis weit über dem Durchschnitt) bewertet. Jeder Fall entspricht einem Geschmackstester.
- **telco.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Telekommunikationsunternehmens geht, die Kundenabwanderung zu verringern. Jeder Fall entspricht einem Kunden und enthält verschiedene Informationen zum demografischen Hintergrund und zur Servicenutzung.
- **telco_extra.sav.** Diese Datendatei ähnelt der Datei *telco.sav*, allerdings wurden die Variablen "tenure" und die Log-transformierten Variablen zu den Kundenausgaben entfernt und durch standardisierte Log-transformierte Variablen ersetzt.
- **telco_missing.sav.** Diese Datendatei ist eine Untermenge der Datendatei *telco.sav*, allerdings wurde ein Teil der demografischen Datenwerte durch fehlende Werte ersetzt.
- **testmarket.sav.** Diese hypothetische Datendatei bezieht sich auf die Pläne einer Fast-Food-Kette, einen neuen Artikel in ihr Menü aufzunehmen. Es gibt drei mögliche Kampagnen zur Verkaufsförderung für das neue Produkt. Daher wird der neue Artikel in Filialen in mehreren zufällig ausgewählten Märkten eingeführt. An jedem Standort wird eine andere Form der Verkaufsförderung verwendet und die wöchentlichen Verkaufszahlen für das neue Produkt werden für die ersten vier Wochen aufgezeichnet. Jeder Fall entspricht einer Standort-Woche.
- **testmarket_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *testmarket.sav*, wobei die wöchentlichen Verkaufszahlen zusammengefasst sind, sodass jeder Fall einem Standort entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und die verzeichneten Verkaufszahlen sind nun die Summe der Verkaufszahlen während der vier Wochen der Studie.
- **tree_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_credit.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält.
- **tree_missing_data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält und eine große Anzahl fehlender Werte aufweist.
- **tree_score_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_textdata.sav.** Eine einfache Datendatei mit nur zwei Variablen, die vor allem den Standardzustand von Variablen vor der Zuweisung von Messniveau und Wertelabels zeigen soll.
- **tv-survey.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Studie, die von einem Fernsehstudio durchgeführt wurde, das überlegt, ob die Laufzeit eines erfolgreichen Programms verlängert werden soll. 906 Personen wurden gefragt, ob sie das

Programm unter verschiedenen Bedingungen ansehen würden. Jede Zeile entspricht einem Befragten; jede Spalte entspricht einer Bedingung.

- **ulcer_recurrence.sav.** Diese Datei enthält Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren. Es stellt ein gutes Beispiel für intervallzensierte Daten dar und wurde an anderer Stelle () vorgestellt und analysiert.
- **ulcer_recurrence_recoded.sav.** In dieser Datei sind die Daten aus *ulcer_recurrence.sav* so umstrukturiert, dass das Modell der Ereigniswahrscheinlichkeit für jedes Intervall der Studie berechnet werden kann und nicht nur die Ereigniswahrscheinlichkeit am Ende der Studie. Sie wurde an anderer Stelle () vorgestellt und analysiert.
- **verd1985.sav.** Diese Datendatei enthält eine Umfrage (). Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die relevanten Variablen sind in drei Sets unterteilt. Set 1 umfasst *alter* und *heirat*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* wird mehrfach nominal skaliert und die Variable *Alter* ordinal. Alle anderen Variablen werden einzeln nominal skaliert.
- **virus.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Internet-Dienstanbieters geht, der die Auswirkungen eines Virus auf seine Netzwerke ermitteln möchte. Dabei wurde vom Moment der Virusentdeckung bis zu dem Zeitpunkt, zu dem die Virusinfektion unter Kontrolle war, der (ungefähre) prozentuale Anteil infizierter E-Mail in den Netzwerken erfasst.
- **waittimes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Wartezeiten für Kunden bei drei verschiedenen Filialen einer Bank. Jeder Fall entspricht einem Kunden und zeichnet die Wartezeit und die Filiale.
- **webusability.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests zur Benutzerfreundlichkeit eines neuen Internetgeschäfts geht. Jeder Fall entspricht einer von fünf Testpersonen, die die Benutzerfreundlichkeit bewerten und gibt für sechs separate Aufgaben an, ob die Testperson sie erfolgreich ausführen könnte.
- **wheeze_steubenville.sav.** Hierbei handelt es sich um eine Teilmenge der Daten aus einer Langzeitstudie zu den gesundheitlichen Auswirkungen der Luftverschmutzung auf Kinder (). Die Daten enthalten wiederholte binäre Messungen des Keuchens von Kindern aus Steubenville, Ohio, im Alter von 7, 8, 9 und 10 Jahren sowie eine unveränderlichen Angabe, ob die Mutter im ersten Jahr der Studie rauchte oder nicht.
- **workprog.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einem Arbeitsprogramm der Regierung, das versucht, benachteiligten Personen bessere Arbeitsplätze zu verschaffen. Eine Stichprobe potenzieller Programmteilnehmer wurde beobachtet. Von diesen Personen wurden nach dem Zufallsprinzip einige für die Teilnahme an dem Programm ausgewählt. Jeder Fall entspricht einem Programmteilnehmer.

Index